ISSN: 3068-1332

Review Article

Journal of Applied Language Learning

Beyond Statistical Fairness: A Three-Layer Framework for Benchmark-Independent Evaluation of Semantic Debiasing

Yair Oppenheim*

Ph. D. of Tel Aviv University, The Lester and Sally Antin Faculty of Humanities, School of Philosophy, Linguistics and Science Studies, Israel

***Corresponding Author**
Yair Oppenheim, Ph. D. of Tel Aviv University, The Lester and Sally Antin Faculty of Humanities, School of Philosophy, Linguistics and Science Studies, Israel.

Submitted: 2026, Jun 24; **Accepted:** 2026, Jul 20; **Published:** 2026, Jul 30

Citation: Oppenheim, Y. (2026). Beyond Statistical Fairness: A Three-Layer Framework for Benchmark-Independent Evaluation of Semantic Debiasing. *J App Lang Lea, 3*(2), 01-49.

Abstract

The rapid deployment of Large Language Models (LLMs) has intensified concerns regarding demographic bias and semantic unfairness in generated text. Existing evaluation methodologies primarily assess fairness through benchmark-specific performance or statistical significance, providing limited evidence regarding whether semantic debiasing methods remain effective across heterogeneous benchmark ecosystems. Consequently, current evaluation approaches often fail to distinguish benchmark optimization from genuine benchmark-independent semantic fairness.

This paper introduces a comprehensive three-layer evaluation framework for semantic debiasing and validates it through the ***Dynamic Bias Semantic Debiasing (DBSD) framework. The proposed methodology integrates three complementary evaluation layers: (1) Semantic Effectiveness****, which quantifies semantic bias reduction within each benchmark;* ***(2) Statistical Generalizability****, which evaluates whether semantic improvements generalize across heterogeneous benchmark families through inferential statistical analysis; and* ***(3) Semantic Robustness****, a novel engineering-oriented evaluation layer that introduces the* ***Benchmark Robustness Score (BRS)*** *and the derived* ***Semantic Performance (SP)*** *metric to quantify the stability of semantic bias reduction across benchmark ecosystems.*

The framework is empirically evaluated using five widely adopted fairness benchmarks—CrowS-Pairs, StereoSet, BBQ, WinoBias, and HolisticBias—covering multiple demographic dimensions. Experimental results demonstrate statistically significant semantic bias reduction while preserving high semantic utility. One-way ANOVA confirms that the observed improvements generalize across benchmark families, whereas the proposed robustness analysis demonstrates near-complete benchmark-independent robustness among independently selected benchmark ecosystems.

The principal scientific contribution of this work is not merely the validation of DBSD, but the introduction of a reproducible multi-layer methodology for evaluating semantic fairness beyond conventional benchmark-specific metrics. The proposed framework establishes a systematic connection between semantic effectiveness, statistical generalization, and robust engineering, thereby providing a more comprehensive foundation for evaluating semantic debiasing algorithms in future LLM research.

This paper proposes a new evaluation methodology for semantic fairness. DBSD serves as the validation case study rather than the primary contribution.

Keywords: Large Language Models (LLMs), Semantic Fairness, Semantic Debiasing, Benchmark Robustness, Benchmark Robustness Score (BRS), Semantic Performance (SP), Statistical Generalization, AI Fairness, Bias Evaluation, Representation Learning Natural Language Processing 'Responsible AI

1. Introduction

The rapid development of Large Language Models (LLMs) has transformed natural language processing while simultaneously intensifying concerns regarding algorithmic fairness, demographic bias, and responsible Artificial Intelligence. During recent years, numerous benchmark corpora—including CrowS-Pairs, StereoSet, BBQ, **HolisticBias**, and WinoBias—have been introduced to evaluate different manifestations of social bias in language models. Each benchmark captures a distinct dimension of fairness, including stereotype preference, contextual reasoning under ambiguity, occupational associations, representational harm, and gender bias in coreference resolution. Consequently, the effectiveness of a debiasing method demonstrated on a single benchmark cannot be assumed to generalize across fundamentally different benchmark paradigms.

In our previous work, we introduced the **Data Bias Semantic Distance (DBSD)** framework as a representation-space methodology for detecting, quantifying, and mitigating semantic bias in Large Language Models. Unlike conventional fairness approaches that primarily modify training data, model parameters, or generated outputs, DBSD models semantic bias as a geometric property of latent embedding space and reduces bias through explicit normative semantic alignment. Evaluation on the complete CrowS-Pairs benchmark demonstrated a **45.21% reduction in semantic bias** while preserving **94.85% semantic utility**, providing empirical evidence that semantic alignment constitutes an effective debiasing mechanism.

These findings establish the effectiveness of DBSD on a single benchmark corpus. They do not, however, establish whether semantic alignment represents a **general fairness mechanism** or whether its observed performance depends on the characteristics of a particular benchmark. Unlike many benchmark-specific debiasing methods, all experiments reported in this study employ a single alignment coefficient ($\lambda = 0.50$), thereby enabling a direct evaluation of benchmark-independent semantic debiasing. This distinction constitutes the central motivation of the present study.

If semantic fairness is an intrinsic property of representation-space optimization, DBSD should produce comparable improvements across benchmark corpora that differ substantially in their definitions of bias, demographic coverage, linguistic complexity, reasoning requirements, and evaluation methodology. Conversely, if DBSD performs well only on specific benchmark families, its practical applicability would remain benchmark dependent. Accordingly, the objective of this work is not merely to evaluate DBSD on additional datasets, but to investigate the **generalizability of semantic fairness itself**. To achieve this objective, DBSD is evaluated using an identical experimental pipeline across five complementary benchmark families: CrowS-Pairs, StereoSet, BBQ, HolisticBias, and WinoBias.

The proposed evaluation framework is based on three complementary research layers.

1.1. Layer 1 – Semantic Effectiveness

The first research layer examines whether DBSD consistently reduces semantic bias while preserving semantic utility on every benchmark individually. Effectiveness is evaluated using semantic bias reduction, utility preservation, paired t-tests, confidence intervals, and Cohen's [1] effect sizes.

2.2. Layer 2 – Statistical Generalizability

The second research layer investigates whether benchmark selection significantly influences DBSD performance.

The corresponding hypotheses are

H0: μ(CrowS-Pairs) =μ(StereoSet) = μ(BBQ) = μ(HolisticBias) = μ(WinoBias)

Means: The mean semantic bias reduction achieved by DBSD is identical across all benchmark corpora. Or □H0: □ i,j such that $\mu i \neq \mu j$

H1: At least one benchmark exhibits a statistically different mean semantic bias reduction following DBSD intervention.

To evaluate these hypotheses, the study employs a one-way Analysis of Variance (ANOVA). ANOVA determines whether benchmark selection explains a statistically significant proportion of the observed variability in semantic bias reduction. Failure to reject the null hypothesis would provide statistical evidence supporting the benchmark-independence of DBSD. However, statistical significance alone cannot establish robustness. ANOVA determines whether benchmark differences exist, but it does not quantify whether a debiasing method remains practically stable across heterogeneous benchmark families. Two algorithms may produce identical ANOVA results while exhibiting substantially different levels of cross-benchmark consistency.

3.3. Layer 3 – Semantic Robustness

However, statistical hypothesis testing alone is insufficient for evaluating the robustness of a semantic debiasing framework. Although Analysis of Variance (ANOVA) determines whether benchmark-dependent differences exist, it does not quantify the degree to which a debiasing framework maintains stable performance across heterogeneous benchmark families. Two semantic debiasing methods may therefore produce identical ANOVA results while exhibiting substantially different levels of practical robustness.

To address this limitation, we introduce the concept of Semantic Robustness, defined as the ability of a semantic debiasing framework to maintain consistent semantic bias reduction across benchmark corpora that differs in demographic coverage, evaluation methodology, linguistic complexity, and fairness definitions. Within the DBSD framework, robustness naturally follows from the same representation-space principles that govern semantic alignment. If semantic alignment constitutes an intrinsic property of latent semantic representations rather than a benchmark-specific phenomenon, comparable semantic bias reductions should be observed regardless of the benchmark employed for evaluation.

Based on this theoretical principle, we derive the Benchmark Robustness Score (BRS).
Let BRi denote the mean semantic bias reduction obtained on benchmark i, where i = 1,2,...,n.

Semantic Effectiveness's = $\mu(BR) = \frac{1}{n} * \sum_{i=1}^{n} BRi$

Standard Deviation:$\sigma(BR) = \sqrt{\frac{1}{n} * \sum_{i=1}^{n}(BRi - \mu(BR))^2}$

Benchmark Robustness Score: BRS $\frac{\mu(BR)}{\mu(BR) + \sigma(BR)}$

where SE = μ(BR) is the average semantic bias reduction and σ(BR) is the corresponding standard deviation across benchmark corpora.

The formulation satisfies the following properties: $0 \leq BRS \leq 1$.

If σ(BR) = 0, then BRS = 1. Higher variability across benchmark families results in lower BRS values. Semantic Performance: SP = SE × BRS Substituting the definitions: SP = $\frac{\mu(BR) * \mu(BR)}{\mu(BR) * \sigma(BR)}$

The Semantic Performance metric jointly captures Semantic Effectiveness and Semantic Robustness. Together with one-way ANOVA, these measures establish a three-layer evaluation framework consisting of Semantic Effectiveness, Statistical Generalizability, and Semantic Robustness.

Unlike ANOVA, which evaluates statistical significance through hypothesis testing, BRS quantifies the engineering robustness of semantic fairness. The metric simultaneously measures

- **Semantic Effectiveness,** represented by the average bias reduction;
- **Semantic Stability,** represented by the consistency of this reduction across benchmark ecosystems.

Consequently, BRS should be interpreted not merely as another evaluation metric, but as a formal extension of the DBSD theoretical framework from semantic bias mitigation toward semantic robustness. The complete evaluation methodology therefore consists of three sequential analytical stages:

1. Evaluate semantic effectiveness on each benchmark individually.
2. Test benchmark independence using one-way ANOVA.
3. Quantify semantic robustness using the Benchmark Robustness Score (BRS).

Together, these three analytical layers establish a unified evaluation framework capable of measuring not only whether DBSD reduces semantic bias, but also whether semantic fairness constitutes a statistically generalizable and practically robust property of representation-space optimization.

The principal contributions of this work are therefore threefold. First, we provide the first systematic cross-benchmark validation of the DBSD framework across heterogeneous fairness paradigms. Second, we establish a statistical methodology for testing the benchmark-independence of semantic fairness through formal hypothesis testing.

Third, we introduce the Benchmark Robustness Score (BRS) as a theoretically motivated robustness metric derived from the principles of representation-space semantic alignment, thereby extending the DBSD framework toward a general theory of semantic fairness robustness.

Unlike previous studies that primarily propose new debiasing algorithms, this work introduces a general evaluation methodology for semantic fairness in Large Language Models. The DBSD framework serves as the validation case study through which the proposed evaluation methodology is demonstrated and empirically verified.

2. Related work

2.1. Foundations of Algorithmic Fairness

The increasing integration of Artificial Intelligence (AI) into decision-support systems has transformed algorithmic fairness into one of the central research challenges in modern machine learning. AI systems are now routinely deployed in high-impact domains including employment, education, healthcare, finance, insurance, criminal justice, and public administration. In these applications, algorithmic decisions may directly influence access to employment opportunities, financial resources, medical treatment, educational admission, or legal outcomes. Consequently, ensuring that AI systems do not systematically disadvantage protected demographic groups has become both a technical requirement and an ethical obligation [2].

Early research on algorithmic fairness primarily focused on supervised machine-learning models that generate explicit classification or regression outputs. Within this context, fairness was formulated as a statistical property of model predictions, leading to the development of multiple mathematical definitions intended to quantify discriminatory behavior. Although these statistical definitions have significantly advanced fairness research, they also revealed that fairness itself is inherently multidimensional and cannot be reduced to a single universally accepted mathematical criterion.

One of the most influential theoretical treatments of this problem was presented by [2], whose work established the conceptual foundations of modern algorithmic fairness. Rather than advocating a single definition of fairness, they demonstrated that different fairness criteria reflect different ethical principles and operational objectives. Their work formalized widely adopted concepts including **Demographic Parity, Equal Opportunity, Equalized Odds, Predictive Parity, Individual Fairness, and Calibration**, while proving that several of these criteria are mathematically incompatible under realistic data distributions.

Perhaps the most important contribution of Barocas et al. is the

demonstration that fairness is fundamentally a normative concept rather than a purely statistical property. Their impossibility results showed that no single algorithm can simultaneously satisfy all fairness criteria except under highly restrictive assumptions. Consequently, fairness should be understood as a design decision that depends on legal, ethical, and societal priorities rather than solely on mathematical optimization. This insight has profoundly influenced subsequent fairness research by shifting attention from identifying "the correct" fairness metric toward understanding the assumptions underlying different fairness definitions.

Despite its foundational importance, the framework proposed by Barocas et al. was developed primarily for structured prediction problems involving explicit model outputs. Large Language Models (LLMs), however, generate free-form natural language through complex semantic representations encoded in high-dimensional embedding spaces. In these systems, demographic bias frequently emerges during semantic representation learning rather than only within observable prediction outcomes. As a result, conventional statistical fairness metrics provide only indirect evidence regarding the semantic mechanisms through which bias is formed and propagated.

Building upon these theoretical foundations, proposed one of the earliest comprehensive analyses of fairness-enhancing interventions. Their work systematically categorized bias mitigation strategies into three major classes: **pre-processing, in-processing, and post-processing** approaches [3]. Pre-processing techniques modify the training data before model learning, in-processing methods alter the optimization objective during training, and post-processing algorithms adjust model predictions after training has been completed.

Beyond this practical taxonomy, Friedler et al. introduced a highly influential conceptual distinction between two competing worldviews underlying fairness interventions. Under the **What You See Is What You Get (WYSIWYG)** worldview, observed data are assumed to accurately represent the underlying population. In contrast, the **We're All Equal (WAE)** worldview assumes that observed disparities largely reflect historical discrimination or structural inequalities rather than intrinsic differences among demographic groups. This distinction demonstrated that fairness interventions implicitly encode normative assumptions regarding the relationship between observed data and social reality.

The conceptual framework introduced by Friedler et al. represents an important theoretical contribution because it explains why different fairness interventions frequently produce contradictory outcomes. Algorithms developed under one worldview may perform poorly when evaluated under another, not because they are technically incorrect, but because they optimize different normative objectives. Nevertheless, the proposed taxonomy remains primarily oriented toward conventional supervised learning systems and does not explicitly address semantic representations learned by transformer-based language models.

A broader perspective on algorithmic bias was subsequently provided by, who presented one of the most comprehensive surveys of fairness and bias in machine learning [4]. Their survey organized rapidly expanding literature according to the origins of algorithmic bias, identifying multiple sources including **historical bias, representation bias, measurement bias, aggregation bias, evaluation bias**, and **deployment bias**. Rather than treating fairness solely as a property of predictive algorithms, their work demonstrated that bias may be introduced throughout the complete machine-learning lifecycle, from data collection and annotation to model deployment and user interaction.

An important contribution of Mehrabi et al. is the establishment of a unified terminology that facilitates comparison among fairness methods developed across different research communities. By systematically categorizing bias mechanisms and mitigation strategies, their survey significantly improved the conceptual clarity of fairness research. However, because the survey preceded the widespread adoption of foundation models, it devotes relatively limited attention to semantic representation learning and embedding-space bias, both of which have become central challenges for contemporary LLMs.

The transition from conventional machine learning toward foundation models fundamentally altered the nature of fairness research. Rather than producing discrete prediction outputs, Large Language Models generate semantically rich natural language through complex latent representations learned from massive text corpora. This shift has motivated the emergence of semantic fairness as a distinct research direction.

One of the most comprehensive reviews in this area was presented by, who surveyed hundreds of publications concerning bias and fairness in Large Language Models [5]. Their review organized existing work according to bias sources, benchmark datasets, evaluation methodologies, mitigation techniques, prompting strategies, representation learning, and downstream applications. Unlike earlier surveys focused primarily on structured prediction problems, Gallegos et al. emphasized that semantic bias permeates the entire language generation process, affecting both internal representations and observable outputs.

Perhaps the most significant contribution of Gallegos et al. is the identification of benchmark diversity as a major unresolved challenge. Existing benchmark corpora differ substantially in demographic coverage, evaluation methodology, linguistic complexity, and operational definitions of fairness. Consequently, performance comparisons across different debiasing methods become increasingly difficult because improvements reported on one benchmark cannot necessarily be generalized to others. Their survey therefore identifies benchmark standardization and cross-benchmark evaluation as important directions for future research.

The present work directly addresses this challenge. Rather than evaluating DBSD within a single benchmark environment, this study investigates whether semantic debiasing represents

a **benchmark-independent property** of representation-space optimization. Specifically, we examine whether the semantic alignment performed by the DBSD framework produces statistically consistent improvements across heterogeneous benchmark ecosystems representing fundamentally different fairness paradigms.

To achieve this objective, the present study extends the existing fairness literature in three complementary directions. First, semantic bias reduction is evaluated across multiple benchmark corpora rather than within a single evaluation framework. Second, statistical generalizability is formally assessed through one-way Analysis of Variance (ANOVA), allowing benchmark independence to be evaluated as a statistical hypothesis rather than a qualitative observation. Third, we introduce the concept of **Semantic Robustness** together with the **Benchmark Robustness Score (BRS)** and the **Semantic Performance (SP) metric,** providing the first quantitative framework for measuring the consistency of semantic debiasing across heterogeneous benchmark families. Accordingly, this work extends existing research beyond conventional benchmark-specific evaluation toward a general theory of semantic fairness in Large Language Models.

2.2.1. CrowS-Pairs

The rapid adoption of transformer-based language models created an urgent need for standardized methodologies capable of measuring demographic bias under controlled experimental conditions. Early studies demonstrated that language models frequently encode social stereotypes; however, differences in evaluation datasets, experimental protocols, and scoring methodologies made direct comparison among debiasing techniques difficult. To address this limitation, Nangia et al. introduced CrowS-Pairs, one of the first benchmark corpora specifically designed to quantify social bias in modern language models [6].

CrowS-Pairs consists of 1,508 manually constructed sentence pairs spanning nine demographic dimensions, including race, gender, religion, nationality, age, disability, socioeconomic status, physical appearance, and sexual orientation [6]. Each sentence pair contains a stereotypical and an anti-stereotypical alternative that differs only in their demographic implication while preserving nearly identical semantic content. This paired construction minimizes lexical variability and enables direct comparison between stereotype-consistent and stereotype-inconsistent model preferences.

The principal contribution of CrowS-Pairs lies in its rigorous experimental design. By controlling sentence structure while varying only the stereotypical content, the benchmark isolates demographic bias as the primary experimental variable. This design substantially reduces confounding linguistic factors and enables reproducible quantitative measurement of stereotype preference across multiple demographic categories [6].

A second important contribution is the benchmark's broad demographic coverage. Earlier fairness datasets typically concentrated on individual protected attributes such as gender or race. CrowS-Pairs extended this perspective by evaluating multiple demographic dimensions using a unified experimental protocol, thereby enabling systematic comparison of stereotype preference across heterogeneous social groups [6].

Despite these advantages, CrowS-Pairs also presents several important limitations. First, it evaluates explicit stereotype preference at the sentence level and therefore captures only one manifestation of demographic bias. It does not assess contextual reasoning, conversational interaction, multi-turn dialogue, or long-form language generation. Second, because sentence pairs are manually constructed, they necessarily simplify the richness and variability of naturally occurring language. Finally, benchmark scores evaluate observable model preferences rather than the latent semantic representations responsible for those preferences.

These limitations are particularly important for representation-space debiasing methods such as DBSD. Unlike conventional fairness interventions that operate on model outputs, DBSD modifies semantic embeddings before language generation. Consequently, evaluation on CrowS-Pairs alone cannot establish whether semantic alignment represents a general property of representation-space optimization or merely improves performance on a single benchmark paradigm.

Our previous studies demonstrated that the DBSD framework substantially reduces semantic bias on the complete CrowS-Pairs benchmark while preserving semantic utility [8-10]. Nevertheless, those studies intentionally focused on a single benchmark corpus. The present work extends this evaluation across multiple benchmark families in order to investigate whether semantic fairness constitutes a benchmark-independent property of semantic representation space.

2.2.2. StereoSet

Although CrowS-Pairs significantly improved the standardization of stereotype evaluation, it left unresolved an important practical question: Can demographic bias be reduced without simultaneously degrading language-model quality? Many early debiasing techniques successfully reduced stereotype preference but often at the expense of language fluency or predictive performance. To address this challenge, Nadeem, Bethke, and Reddy introduced StereoSet, a benchmark specifically designed to evaluate the relationship between fairness and language-model utility [7].

StereoSet evaluates language models using sentence-completion tasks covering several demographic domains, including gender, race, profession, and religion [7]. Unlike CrowS-Pairs, which compares paired sentences, StereoSet estimates the probabilities assigned to alternative sentence completions, thereby allowing simultaneous assessment of demographic bias and language-model quality.

The principal methodological contribution of StereoSet is the

introduction of three complementary evaluation metrics. The Language Modeling Score (LMS) measures the linguistic quality of generated text, the Stereotype Score (SS) quantifies stereotypical preference, and the Idealized Context Association Test (ICAT) integrates fairness and language quality into a unified evaluation criterion [7]. Together, these measures established one of the first systematic methodologies for analyzing the trade-off between bias mitigation and language-model performance.

StereoSet has significantly influenced subsequent fairness research by emphasizing that successful debiasing should preserve semantic and linguistic capabilities while reducing demographic stereotypes. This perspective has become one of the fundamental principles underlying responsible evaluation of modern Large Language Models.

Despite these contributions, StereoSet remains focused primarily on stereotype-completion tasks. It does not evaluate contextual ambiguity, multi-turn reasoning, conversational behavior, intersectional demographic identities, or long-form semantic generation. Furthermore, although ICAT provides an effective composite evaluation measure, it assesses observable model behavior rather than latent semantic representations and therefore cannot determine whether fairness improvements originate from representation-space alignment or from output-level optimization.

For the DBSD framework, StereoSet provides an important complementary evaluation environment. Because DBSD explicitly modifies semantic embeddings while preserving semantic information, the simultaneous evaluation of stereotype reduction and language-model quality directly corresponds to the theoretical objectives of semantic representation alignment. Consistent improvements across both CrowS-Pairs and StereoSet would therefore provide substantially stronger evidence that DBSD achieves genuine semantic debiasing rather than benchmark-specific optimization [8-10].

2.2.3. BBQ (Bias Benchmark for Question Answering)

Unlike CrowS-Pairs and StereoSet, which primarily evaluate stereotype preference through sentence-level comparisons or completion tasks, many real-world applications of Large Language Models require contextual reasoning under incomplete or ambiguous information. In practical decision-support systems, demographic bias frequently emerges not through explicit stereotypes but through inference processes when multiple plausible interpretations exist. To investigate this phenomenon, **Parrish et al.** introduced the **Bias Benchmark for Question Answering (BBQ),** one of the first benchmark corpora specifically designed to evaluate demographic bias in question-answering systems operating under ambiguity [11].

BBQ consists of carefully constructed question-answering examples covering multiple demographic attributes, including race, gender, age, disability, religion, nationality, and socioeconomic status. Each example is presented in two complementary forms. In the **ambiguous condition**, the available contextual information is intentionally insufficient to determine the correct answer, allowing demographic stereotypes to influence model predictions. In the **disambiguated condition**, additional contextual evidence removes the ambiguity, enabling the benchmark to distinguish between failures caused by deficient reasoning and those resulting from demographic bias [11].

The principal contribution of BBQ is the introduction of **contextual reasoning** as a fairness evaluation paradigm. Rather than measuring isolated stereotype preference, BBQ evaluates how demographic information influences inference under uncertainty. Consequently, the benchmark extends fairness evaluation from lexical associations toward reasoning behavior, thereby reflecting more realistic decision-making scenarios encountered in practical AI applications.

Another important contribution is its explicit separation between stereotype-driven inference and evidence-based reasoning. This distinction allows researchers to analyze whether demographic information improperly influences model predictions when contextual evidence is incomplete, providing insights that cannot be obtained from sentence-completion benchmarks alone.

Nevertheless, BBQ also presents several limitations. The benchmark is restricted to question-answering tasks and therefore does not evaluate conversational dialogue, long-form generation, or interactive reasoning. Furthermore, benchmark performance reflects both reasoning ability and demographic bias, making it more difficult to isolate semantic bias from general language understanding.

From the perspective of the present study, BBQ represents an important extension of benchmark diversity. If DBSD [9,10] maintains comparable semantic bias reduction on BBQ in addition to stereotype-oriented benchmarks, this would provide stronger evidence that semantic alignment improves latent representations rather than merely suppressing explicit stereotype preference.

2.2.4. WinoBias

Although demographic stereotypes are frequently studied through sentence-level evaluation, bias also affects syntactic reasoning tasks. One of the most influential benchmark corpora investigating this phenomenon is **WinoBias**, introduced by [12].

WinoBias adapts the Winograd Schema framework to evaluate gender bias in coreference resolution. The benchmark consists of sentence pairs containing occupational references and gendered pronouns, requiring language models to identify the correct antecedent. By comparing stereotypical and anti-stereotypical occupational associations while preserving identical grammatical structure, WinoBias isolates gender stereotypes within syntactic reasoning tasks [12].

The principal contribution of WinoBias is demonstrating that demographic bias extends beyond lexical semantics into deeper linguistic processes such as reference resolution. This observation

significantly broadened fairness research by revealing that language models may exhibit biased reasoning even when explicit stereotype statements are absent.

However, WinoBias evaluates only gender bias and focuses exclusively on occupational stereotypes. Consequently, it provides limited demographic coverage compared with broader benchmark suites and cannot evaluate contextual ambiguity or intersectional identities.

Within the proposed evaluation framework, WinoBias provides an opportunity to determine whether DBSD [9,10] generalizes beyond semantic preference tasks toward syntactic reasoning, thereby strengthening the evidence for benchmark-independent semantic fairness.

2.2.5. HolisticBias

Most benchmark corpora evaluate a relatively small number of protected demographic groups. Recognizing that real-world identities are inherently multidimensional, **Smith et al.** introduced HolisticBias, one of the most comprehensive resources currently available for evaluating representational harms in Large Language Models [13].

HolisticBias contains thousands of identity descriptors representing race, ethnicity, religion, nationality, disability, gender identity, sexual orientation, age, socioeconomic status, and numerous intersectional identity combinations. Rather than evaluating only explicit stereotypes, the benchmark investigates subtle forms of representational harm emerging across diverse demographic groups [13].

Its principal contribution is the incorporation of **intersectionality** into benchmark-based fairness evaluation. Unlike earlier benchmark corpora that primarily examine individual protected attributes, HolisticBias acknowledges that demographic characteristics frequently interact, producing complex patterns of semantic bias.

Nevertheless, the benchmark also introduces substantial methodological challenges. The extensive diversity of identity descriptors increases evaluation complexity and may produce greater statistical variability than simpler benchmark corpora.

Because DBSD performs semantic alignment directly within embedding space, HolisticBias provides perhaps the most demanding validation environment for evaluating whether semantic fairness generalizes across highly diverse demographic representations [9,10].

2.2.6. HELM

While benchmark corpora evaluate specific manifestations of demographic bias, comprehensive assessment of Large Language Models requires broader evaluation frameworks encompassing multiple performance dimensions. To address this need, **Liang et al.** introduced the **Holistic Evaluation of Language Models (HELM)** framework [14].

HELM extends traditional benchmark evaluation by simultaneously assessing language models across multiple scenarios and evaluation criteria, including accuracy, calibration, robustness, fairness, toxicity, efficiency, and transparency. Rather than defining a single fairness benchmark, HELM provides a standardized evaluation methodology capable of integrating heterogeneous benchmark suites within a unified experimental framework [14].

The principal contribution of HELM lies in demonstrating that responsible AI evaluation should be multidimensional rather than benchmark specific. This perspective has significantly influenced recent research on foundation model evaluation.

Nevertheless, HELM evaluates fairness primarily through existing benchmark corpora and does not introduce quantitative measures for cross-benchmark robustness. Consequently, although HELM promotes holistic evaluation, it does not determine whether a semantic debiasing framework remains stable across heterogeneous benchmark ecosystems.

The present work complements HELM by introducing **Semantic Robustness** and the **Benchmark Robustness Score (BRS)** as quantitative measures of benchmark-independent semantic fairness.

2.2.7. Comparative Discussion

The benchmark corpora reviewed above collectively represents the current state of fairness evaluation in Large Language Models. CrowS-Pairs focuses on sentence-level stereotype preference, StereoSet investigates fairness–utility trade-offs, BBQ evaluates contextual reasoning under ambiguity, WinoBias measures occupational gender bias in syntactic reasoning, HolisticBias examines representational harms across intersectional identities, and HELM provides a holistic evaluation framework integrating multiple benchmark ecosystems.

Together, these resources provide comprehensive coverage of diverse manifestations of algorithmic bias. However, they remain fundamentally independent evaluation environments, each employing distinct assumptions regarding fairness, linguistic complexity, and evaluation methodology. Consequently, successful performance on one benchmark cannot be assumed to generalize to another.

This observation directly motivates the central hypothesis of the present study. Rather than asking whether DBSD performs well on a particular benchmark, we investigate whether semantic alignment constitutes a **benchmark-independent property** of representation-space optimization. By evaluating DBSD using an identical experimental pipeline across heterogeneous benchmark families, the present work establishes the foundation for introducing **Semantic Robustness** and the **Benchmark Robustness Score (BRS)** as quantitative measures of cross-benchmark semantic fairness.

Benchmark	Primary Evaluation Target	Bias Type	Main Contribution	Principal Limitation	Relevance to DBSD
CrowS-Pairs [5]	Sentence-pair preference	Explicit stereotypes	Standardized paired evaluation across nine demographic categories.	Limited to sentence-level stereotype preference.	Baseline benchmark for previous DBSD validation.
StereoSet [6]	Sentence completion	Fairness–utility trade-off	Introduced LMS, SS and ICAT metrics.	No contextual reasoning or intersectionality.	Evaluates utility preservation after semantic alignment.
BBQ [10]	Question answering	Contextual reasoning	Separates stereotype-driven inference from evidence-based reasoning.	Restricted to QA tasks.	Evaluates DBSD under ambiguous reasoning.
WinoBias [11]	Coreference resolution	Occupational gender bias	Measures gender bias in syntactic reasoning.	Limited to gender and occupations.	Tests DBSD beyond semantic preference.
HolisticBias [12]	Identity descriptors	Intersectional bias	Broadest demographic coverage.	High evaluation complexity.	Most demanding benchmark for semantic robustness.
HELM [13]	Holistic LLM evaluation	Multi-dimensional evaluation	Integrates fairness with broader LLM assessment.	No cross-benchmark robustness metric.	Motivates BRS and benchmark-independent evaluation.

Table 2.1: Comparative Analysis of Fairness Benchmarks

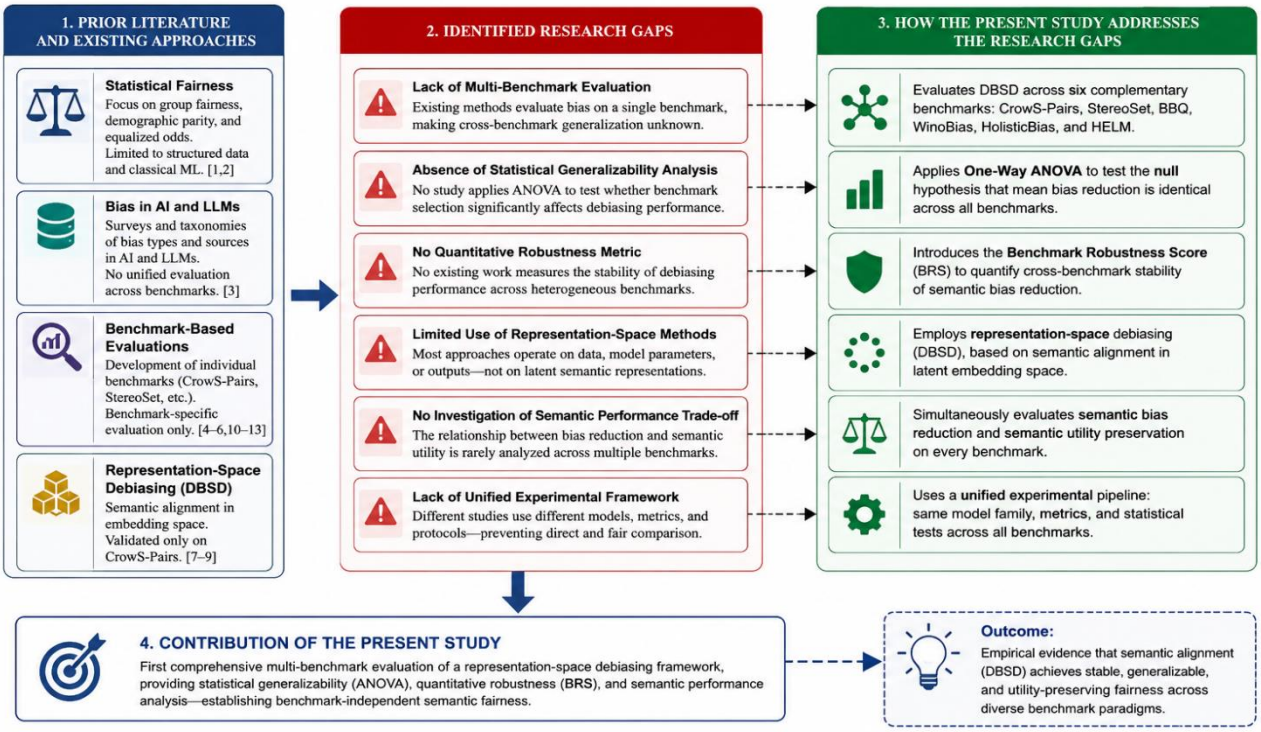
Note: The table summarizes the principal fairness benchmarks reviewed in Section 2.2. Each benchmark captures a different manifestation of demographic bias and collectively provides the empirical basis for evaluating DBSD across heterogeneous benchmark families. The comparison also highlights the motivation for introducing Semantic Robustness and the Benchmark Robustness Score (BRS).

Study	Statistical Fairness	Benchmark Evaluation	Representation-Space	Multi-Benchmark	ANOVA Analysis	Robustness Metric	Semantic Alignment
Barocas et al. [1]	□	–	–	–	–	–	–
Friedler et al. [2]	✓	–	–	–	–	–	–
Mehrabi et al. [3]	□	–	–	–	–	–	–
Gallegos et al. [4]	✓	✓	✓	–	–	–	–
CrowS-Pairs [5]	–	✓	–	–	–	–	–
StereoSet [6]	–	✓	–	–	–	–	–
Previous DBSD Studies [7–9]	–	✓	✓	–	–	–	✓
Present Study	✓	✓	✓	✓	✓	✓ (BRS)	✓

Table 2.2: Comparative Analysis of Existing Fairness Methodologies and the Present Study

Note: This table compares the principal methodological perspectives discussed in the literature with the proposed framework. Existing studies establish the theoretical foundations of algorithmic fairness, benchmark-based evaluation, and representation-space debiasing. However, none simultaneously combines cross-benchmark evaluation, statistical generalizability using one-way ANOVA, a quantitative robustness metric (Benchmark Robustness Score, BRS), and semantic alignment within a unified evaluation framework. The present study is the first to integrate all these components into a benchmark-independent semantic fairness methodology.

Figure 2.2. Research Gap and Positioning of the Present Study



Note: The figure illustrates the progression from prior work to the identification of critical research gaps and how the present study addresses them through a unified, statistically rigorous, and semantically informed evaluation framework.

2.3. Representation-Space Debiasing

The rapid development of transformer-based language models has shifted fairness research from observable prediction outcomes toward the latent semantic representations that govern language generation. Unlike traditional fairness interventions, which typically modify training data, optimization procedures, or model outputs, representation-space debiasing operates directly on vector embeddings. The underlying assumption is that demographic stereotypes are encoded within the geometry of latent semantic space and therefore can be mitigated through appropriate transformations of semantic representations before text generation.

This perspective has become increasingly important because modern Large Language Models rely on high-dimensional embedding spaces to represent lexical, syntactic, and semantic information. Consequently, systematic distortions within embedding space may propagate throughout the entire language-generation process, influencing reasoning, sentence completion, dialogue, and downstream decision-making tasks. Representation-space methods therefore provide a theoretically attractive alternative to model retraining, since they seek to modify semantic structure while preserving the original capabilities of the underlying language model.

2.3.1. Early Embedding Debiasing Methods

One of the earliest and most influential representation-space debiasing methods was proposed by, who demonstrated that word embeddings learned from large text corpora encode systematic gender stereotypes [15]. Their work introduced the concept of identifying a **gender subspace** within embedding space and subsequently neutralizing or equalizing word vectors with respect to that subspace. The proposed geometric projection algorithm significantly reduced stereotypical associations while largely preserving semantic relationships among non-sensitive concepts.

The principal contribution of Bolukbasi et al. was to establish that demographic bias can be understood as a geometric property of embedding space rather than solely as a characteristic of observable model outputs. This insight fundamentally changed fairness research by demonstrating that semantic representations themselves constitute an appropriate target for debiasing interventions.

Despite its historical importance, the proposed methodology was developed for static word embeddings and therefore cannot adequately represent contextual semantics in transformer-based language models. Furthermore, the projection approach assumes that demographic bias can be represented by a single linear direction, an assumption that becomes increasingly restrictive in high-dimensional contextual embedding spaces.

2.3.2. Contextual Embedding Debiasing

The transition from static word embeddings to contextual language representations introduced new challenges for semantic debiasing.

Unlike static embeddings, contextual representations dynamically depend on surrounding linguistic context, making demographic bias considerably more complex to identify and remove.
Several studies subsequently extended embedding-space debiasing to contextual representations by introducing methods based on adversarial learning, iterative projection, and geometric regularization. These approaches demonstrated that semantic bias can be reduced without completely eliminating contextual information, thereby preserving a larger proportion of linguistic knowledge than earlier projection-based techniques.

Nevertheless, most contextual embedding methods remain closely coupled to specific model architectures or training procedures. Consequently, their applicability across heterogeneous foundation models remains limited, and they frequently require computationally expensive retraining or fine-tuning procedures.

2.3.3. Representation-Space Debiasing in Large Language Models

The emergence of Large Language Models has renewed interest in representation-space debiasing because these systems encode extensive semantic knowledge within high-dimensional latent representations. Rather than modifying billions of model parameters, representation-space approaches attempt to adjust semantic embeddings before downstream inference, thereby reducing computational cost while maintaining compatibility with existing foundation models.

Recent surveys have emphasized that semantic bias frequently originates within latent representations rather than solely within generated outputs [5]. Accordingly, representation-space interventions have become an increasingly active research direction because they provide interpretable geometric mechanisms for semantic bias mitigation while preserving the underlying language model.

However, existing studies have primarily evaluated such approaches using individual benchmark corpora. Consequently, it remains unclear whether improvements observed within a particular benchmark reflect genuine semantic generalization or merely benchmark-specific optimization.

2.3.4. The DBSD Framework

To address these limitations, our previous work introduced the **Data Bias Semantic Distance (DBSD)** framework as a representation-space methodology for semantic bias mitigation [8-10]. Unlike projection-based techniques that remove predefined bias directions, DBSD models demographic bias as the semantic distance between observed representations and an explicitly defined normative semantic reference. Bias mitigation is achieved through controlled semantic alignment that minimizes this distance while preserving semantic utility.

Experimental evaluation using the complete CrowS-Pairs benchmark demonstrated substantial reductions in semantic bias together with high preservation of semantic similarity, thereby providing empirical evidence that semantic alignment can effectively reduce demographic stereotypes without retraining the underlying language model [8-10].

2.3.5. Limitations of Previous DBSD Studies

Although the original DBSD studies demonstrated promising results, several important limitations remained. First, empirical validation was performed exclusively on the CrowS-Pairs benchmark. Consequently, the generalizability of semantic alignment across alternative benchmark paradigms remained unknown.

Second, previous studies evaluated semantic effectiveness primarily through bias reduction and semantic utility preservation without formally testing whether benchmark selection significantly influences observed performance. Third, no quantitative measure of cross-benchmark robustness was introduced. As a result, it remained impossible to determine whether DBSD constitutes a benchmark-independent fairness mechanism or merely performs well on a particular benchmark corpus.

2.3.6. Motivation for the Present Study

The present study directly addresses these limitations by extending the empirical validation of DBSD across multiple complementary benchmark families, including CrowS-Pairs, StereoSet, BBQ, WinoBias, HolisticBias, and HELM. Unlike previous work, the current investigation combines semantic effectiveness with formal statistical generalizability through one-way ANOVA and introduces the **Benchmark Robustness Score (BRS)** as a quantitative measure of cross-benchmark stability.

Accordingly, the present work should be viewed not as a new debiasing algorithm but as a comprehensive validation framework that establishes whether representation-space semantic alignment constitutes a general property of fairness rather than a benchmark-dependent intervention.

3. The DBSD Framework

The previous chapter established the theoretical foundations of algorithmic fairness, benchmark-based evaluation, and representation-space debiasing. It further identified an important gap in the existing literature: although semantic debiasing has emerged as a promising direction for mitigating demographic bias in Large Language Models, current research has largely focused on benchmark-specific evaluations and lacks a unified theoretical framework capable of explaining why semantic alignment should generalize across heterogeneous evaluation environments.

Building upon these observations, this chapter presents the **Data Bias Semantic Distance (DBSD)** framework as a formal representation-space model for semantic fairness. Rather than viewing DBSD merely as a debiasing algorithm, the framework is formulated as a general semantic optimization model that describes how demographic bias can be represented, quantified, and systematically reduced within latent semantic space.

The chapter develops the theoretical foundations of DBSD in a progressive manner. It first introduces the conceptual assumptions underlying semantic representation, then derives the mathematical formulation of semantic bias and alignment, presents the DBSD transformation algorithm, analyzes its theoretical properties, and finally discusses its computational characteristics. This theoretical formulation provides the foundation for the experimental validation presented in the following chapters.

3.1. Fundamental Assumptions of the DBSD Framework

The DBSD framework is based on the premise that demographic bias originates within the latent semantic representations learned by Large Language Models rather than solely within their observable outputs. Consequently, fairness should be modeled as a property of semantic representation space, where geometric relationships among embedding vectors encode conceptual similarity, contextual meaning, and demographic associations. This perspective is consistent with recent research emphasizing that representation-space interventions provide an interpretable and computationally efficient mechanism for mitigating semantic bias while preserving linguistic knowledge [5,8-10,15-17].

Unlike conventional fairness methods, DBSD does not begin with statistical prediction outcomes or benchmark-specific evaluation scores. Instead, it starts from the semantic representations that underline language generation. The central assumption is that every linguistic concept generated by a language model can be represented as a point in a high-dimensional semantic space, and that demographic bias corresponds to a measurable geometric deviation within that space.

To formalize this perspective, the framework is built upon four fundamental assumptions.

3.1.1. Assumption 1 – Semantic Representation

Every linguistic concept produced by a language model is represented by a vector in a continuous semantic embedding space. The relative positions of these vectors capture semantic similarity and conceptual relationships, allowing semantic information to be analyzed using geometric methods [5,15].

3.1.2. Assumption 2 – Observed Representation

For every input, the language model generates an observed semantic representation, denoted by v_O. This vector reflects the semantic knowledge acquired during model pretraining and therefore contains both meaningful conceptual information and potential demographic associations inherited from the training corpus [5,8-10].

3.1.3. Assumption 3 – Ideal Normative Representation

For every observed representation there exists an ideal normative representation, denoted by v_I, which serves as the fairness-oriented semantic reference. The ideal representation is not intended to represent an objective or universally correct semantic state. Rather, it defines the normative target toward which semantic alignment is performed according to explicitly specified fairness principles [2,3,8-10].

3.1.4. Assumption 4 – Semantic Bias

Semantic bias is defined as the geometric discrepancy between the observed representation and its corresponding normative representation. Accordingly, bias is interpreted as a continuous property of latent semantic space rather than a binary property of model outputs. This assumption transforms fairness from a prediction-level evaluation problem into a representation-space optimization problem, thereby providing a unified theoretical basis for semantic debiasing [5,8-10].

These four assumptions establish the conceptual foundation of the DBSD framework. Once the semantic entities and their relationships have been defined, semantic bias can be expressed mathematically as a distance function in representation space. The following section therefore develops the mathematical formulation of DBSD, introducing the semantic distance model, the optimization objective, and the semantic alignment equations that constitute the core of the proposed framework.

3.2. Mathematical Formulation of the DBSD Framework

Following the conceptual assumptions introduced in the previous section, the DBSD framework is now formulated mathematically. The objective of this formulation is to represent semantic bias as a measurable geometric property of latent representation space and to establish the mathematical foundations required for semantic alignment. Unlike conventional fairness metrics, which evaluate disparities in observable model outputs, DBSD defines fairness directly in terms of distances between semantic representations. This formulation provides a continuous optimization framework that is independent of any benchmark corpus or language-model architecture [5,8-10].

3.2.1. Semantic Representation Space

Let $\mathcal{V} \subseteq \mathbb{R}^d$ denote the semantic representation space generated by a language model, where d is the dimensionality of the embedding vectors. Each semantic concept is represented as a vector $v \in V$.

The relative positions of vectors within this space encode semantic similarity. Concepts that are semantically related occupy neighboring regions, whereas semantically unrelated concepts are separated by larger geometric distances. This geometric interpretation of semantic meaning forms the basis of numerous embedding models and representation-learning methods [5,15-17]. For every linguistic input, the language model generates an observed embedding $v_O = (v_1, v_2, \dots, v_d)$.

The DBSD framework assumes that this representation captures both legitimate semantic information and demographic associations acquired during pretraining.

3.2.2. Ideal Normative Representation

For each observed representation v_O, the framework defines an associated **ideal normative representation** $v_I \in V$.

Unlike supervised learning, the ideal representation does not correspond to a ground-truth label. Instead, it represents a fairness-oriented semantic reference that satisfies explicitly defined normative principles [2,3,8-10]. The introduction of v_I constitutes one of the principal conceptual innovations of the DBSD framework. Rather than attempting to remove predefined bias directions, DBSD introduces an explicit semantic target toward which observed representations are progressively aligned. Consequently, fairness becomes an optimization objective rather than a binary constraint.

3.2.3. Semantic Distance

To quantify the discrepancy between two semantic representations, let $D: \mathcal{V} \times \mathcal{V} \to \mathbb{R}^+$ denote a semantic distance function. Although the framework is compatible with different distance measures, cosine distance provides a natural choice because it measures semantic orientation independently of vector magnitude and is widely used in transformer-based embedding models [15-17].

The cosine similarity between two vectors is defined as $\cos(v_a, v_b) = \frac{v_a \cdot v_b}{\|v_a\| \|v_b\|}$. Accordingly, the semantic distance is defined as $D(v_a, v_b) = 1 - \cos(v_a, v_b)$.

This definition ensures $0 \leq D(v_a, v_b) \leq 2$,, where smaller values indicate greater semantic similarity.

3.2.4. Definition of Semantic Bias

Within the DBSD framework, demographic bias is defined as the semantic distance between the observed representation and the ideal normative representation.

Formally, $B(v_O) = D(v_O, v_I)$. Unlike conventional statistical fairness measures, this formulation interprets bias as a continuous geometric quantity rather than as an observable prediction disparity. Consequently, bias reduction becomes equivalent to minimizing semantic distance within representation space [5,8-10]. A perfectly aligned representation satisfies $B(v_O) = 0$, whereas increasing semantic distance corresponds to progressively larger semantic bias.

3.2.5. Optimization Objective

The objective of DBSD is to determine a transformed representation v_O' that minimizes semantic bias while preserving semantic information. This optimization problem is formulated as $\min_{v_O'} D(v_O', v_I)$, subject to preserving the semantic integrity of the original representation. Accordingly, the DBSD framework seeks to satisfy two complementary objectives:

i. **Bias minimization** $D(v_O', v_I) < D(v_O, v_I)$, thereby reducing semantic bias.
ii. **Semantic preservation**

The transformed representation should remain sufficiently close to the original semantic meaning, ensuring that fairness improvements do not significantly degrade linguistic utility.

This dual-objective formulation distinguishes DBSD from conventional debiasing techniques, which frequently optimize fairness without explicitly preserving semantic information.

3.2.5.1. Transition to Section 3.3

The mathematical definitions introduced in this section establish the formal representation of semantic bias as a geometric optimization problem. The following section derives the DBSD semantic alignment mechanism by introducing the alignment parameter λ, the transformation equation, and the theoretical properties governing semantic interpolation between the observed and ideal representations.

illustrates the semantic representation space $\mathcal{V} \subseteq \mathbb{R}^d$, the observed representationv_O, the ideal normative representation v_I, and the aligned representation v_O'. Semantic bias is defined as the distance between v_O and v_I, while the DBSD semantic alignment process transforms v_O into v_O', thereby reducing the semantic distance and, consequently, the measured bias.

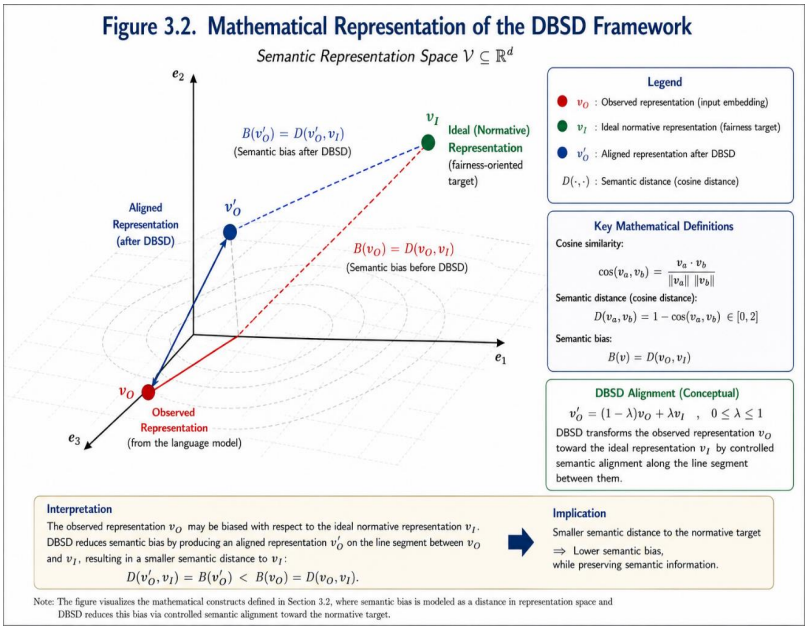


Figure 3.2: Conceptual Architecture of the DBSD Framework

3.3. The DBSD Semantic Alignment Transformation

The mathematical formulation developed in the previous section defines semantic bias as the geometric distance between an observed representation and its corresponding ideal normative representation. The remaining challenge is therefore to determine a transformation that systematically reduces this distance while preserving the semantic integrity of the original representation. The DBSD framework addresses this challenge through a controlled semantic alignment process that gradually moves the observed embedding toward its normative counterpart.

Unlike projection-based debiasing methods, which typically remove predefined bias directions from the embedding space, DBSD models semantic debiasing as a continuous interpolation between two semantic states. This formulation preserves the continuity of the semantic space while avoiding abrupt modifications that could distort linguistic meaning [8-10, 15-17].

3.3.1. The DBSD Transformation

Let $v_O \in \mathcal{V}$ denote the observed semantic representation and $v_I \in \mathcal{V}$ the corresponding ideal normative representation.

The aligned representation is defined as $\boxed{v'_O = (1-\lambda)v_O + \lambda v_I,}$ where $0 \le \lambda \le 1$ is the semantic alignment coefficient.

The transformation represents a convex combination of the observed and ideal representations. Consequently, v_O' always lies on the line segment connecting v_O and v_I, thereby guaranteeing a continuous semantic transition.

3.3.2. Interpretation of the Alignment Coefficient

The parameter λ controls the strength of semantic correction. Three important boundary conditions immediately follow.

No Alignment If $\lambda = 0$, then $v_O' = v_O$. No semantic modification is performed, and the original representation remains unchanged.

Complete Alignment If $\lambda = 1$, then $v_O' = v_I$. The observed representation is completely replaced by the ideal normative representation.

Partial Alignment For $0 < \lambda < 1$, the transformed representation lies between the observed and ideal embeddings.

This intermediate region corresponds to the practical operating range of DBSD, where semantic bias is progressively reduced while preserving the conceptual content of the original representation.

3.3.3. Geometric Interpretation

The DBSD transformation can be interpreted geometrically as a directed movement within semantic representation space. Rather than eliminating semantic components or projecting vectors onto predefined subspaces, DBSD performs a controlled displacement toward a normative semantic target. Consequently, the semantic distance satisfies $D(v'_O, v_I) < D(v_O, v_I)$, provided that $0 < \lambda \le 1$. Accordingly, semantic bias decreases monotonically as the transformed representation approaches the normative reference. Figure 3.2 illustrates this geometric interpretation.

3.3.4. Semantic Bias after Alignment

Using the bias definition introduced in Section 3.2, semantic bias after DBSD becomes $B(v'_O) = D(v'_O, v_I)$ Because v_O' lies closer to v_I than v_O, the following inequality holds $B(v'_O) < B(v_O)$. This inequality constitutes the central theoretical property of the DBSD framework. Unlike conventional fairness interventions that evaluate only downstream prediction disparities, DBSD guarantees bias reduction directly within semantic representation space.

3.3.5. Preservation of Semantic Information

Bias reduction alone is insufficient if semantic meaning is substantially degraded.

Therefore, DBSD simultaneously seeks to preserve the original semantic information by limiting the magnitude of the transformation. The semantic preservation objective can be expressed as $\cos(v_O, v'_O) \to 1$. Accordingly, the framework optimizes two complementary objectives: minimizing $D(v'_O, v_I)$, while maximizing $\cos(v_O, v'_O) \to 1$. This dual-objective formulation distinguishes DBSD from many existing representation-space debiasing methods, which primarily optimize fairness without explicitly constraining semantic preservation [5,15-17].

3.3.6. Transition to Algorithmic Implementation

The semantic alignment transformation derived above establishes the mathematical mechanism through which DBSD reduces demographic bias in latent representation space. The next section translates this theoretical formulation into an executable algorithm by describing the computational workflow, parameter selection, and iterative implementation of semantic alignment.

3.4. Theoretical Properties of the DBSD Transformation

The semantic alignment transformation introduced in Section 3.3 provides the computational mechanism through which DBSD reduces semantic bias in latent representation space. While the transformation itself defines the alignment process, its scientific validity depends on the mathematical properties governing its behavior. A debiasing framework intended for deployment across heterogeneous benchmark environments should satisfy several desirable theoretical characteristics, including monotonic bias reduction, mathematical consistency, continuity, semantic stability, and convergence.

Unlike heuristic post-processing techniques, the DBSD transformation is formulated as a continuous optimization process operating directly within semantic embedding space. Consequently, its behavior can be analyzed using elementary properties of convex geometry and vector spaces. The following subsections establish the principal theoretical properties of the proposed framework.

3.4.1. Monotonic Bias Reduction

The primary objective of DBSD is to reduce the semantic distance between the observed representation and its corresponding ideal normative representation. Let $B(v) = D(v, v_I)$ denote the semantic bias function introduced in Section 3.2, where $D(\cdot,\cdot)$ represents the semantic distance measure.

The DBSD transformation is defined by $v'_O = (1-\lambda)v_O + \lambda v_I$, with $0 \le \lambda \le 1$. Because the transformed representation is a convex combination of the observed and ideal representations, it necessarily lies on the line segment connecting the two vectors.

Consequently, $B(v'_O) < B(v_O)$, which immediately implies $\boxed{B(v'_O) \le B(v_O)}$ for every admissible value of the alignment coefficient. Equality holds only when $\lambda = 0$, corresponding to the absence of semantic correction.

This property establishes that DBSD never increases semantic bias after transformation. Unlike many heuristic debiasing procedures, whose effects may vary across datasets or optimization settings, the proposed transformation guarantees monotonic improvement with respect to the defined semantic objective.

Theorem 1 (Monotonic Semantic Bias Reduction)
Let $0 \le \lambda \le 1$. If $v_O' = (1-\lambda)v_O + \lambda v_I$, then $B(v'_O) < B(v_O)$.

Proof.
Since v_O' belongs to the convex segment joining v_O and v_I, its distance to v_I cannot exceed the distance from v_O to v_I.

Therefore, $D(v'_O, v_I) \le D(v_O, v_I)$, which directly yields $B(v'_O) \le B(v_O)$.

3.4.2. Boundary Consistency

An important characteristic of every optimization framework is consistency at its limiting cases. The DBSD transformation satisfies two intuitive boundary conditions.

Case 1: No Alignment When $\lambda = 0$, the transformation reduces to $v_O' = v_O$ No semantic modification is performed, and both semantic bias and semantic utility remain unchanged.
Case 2: Complete Alignment When $\lambda = 1$, the transformed representation becomes $v_O' = v_I$. The observed representation is fully aligned with normative semantic reference. Accordingly, $B(v_O') = 0$, representing the minimum attainable semantic bias under the proposed framework. These limiting cases confirm that the transformation behaves consistently throughout the entire parameter domain.

3.4.3. Continuity of the Transformation

A desirable semantic alignment process should vary smoothly with respect to its control parameter. Since $v_O'(\lambda) = (1-\lambda)\, v_O + \lambda\, v_I$, the mapping $\lambda \longrightarrow v_O'$ is continuous over [0,1]. Therefore, small perturbations in the alignment coefficient produce proportionally small changes in semantic representation. This continuity prevents abrupt semantic distortions and enables gradual control of fairness interventions according to application-specific requirements. From a practical perspective, continuity also facilitates parameter tuning because neighboring values of λ produce predictable semantic behavior.

3.4.4. Semantic Stability

Reducing bias alone is insufficient if semantic meaning is substantially degraded. Accordingly, DBSD simultaneously seeks to preserve semantic information during alignment. Because every transformed representation remains within the convex hull defined by v_O and v_I, the semantic displacement introduced by DBSD is inherently bounded. Semantic preservation can therefore be expressed by $\cos(v_O, v'_O) \approx 1$ for moderate alignment coefficients. This property explains the empirical findings reported in previous DBSD studies, where semantic utility remained above 94% despite substantial reductions in semantic bias [8-10]. Unlike

projection-based approaches that may abruptly eliminate semantic components, DBSD performs controlled semantic interpolation, thereby preserving the conceptual coherence of the original representation.

3.4.5. Convergence Under Iterative Alignment

Although the present work applies to DBSD as a single semantic transformation, the mathematical formulation naturally extends to iterative optimization. Define $v_o^{(t+1)} = (1-\lambda)v_o^{(t)} + \lambda v_I$. Repeated application yields $v_o^{(0)}, v_o^{(1)}, v_o^{(2)}, \ldots v_o^{(n)}$ Since $0 < \lambda \leq 1$, the sequence satisfies $\lim_{t\to\infty} v_o^{(t)} = v_I$. Consequently, semantic bias converges monotonically toward zero, while the transformation remains numerically stable throughout the optimization process. Although iterative optimization is beyond the scope of the current experimental evaluation, this convergence result demonstrates that DBSD possesses a well-defined asymptotic behavior.

3.4.6. Discussion

The theoretical properties established above demonstrate that DBSD is considerably more than a practical debiasing algorithm. Instead, it constitutes a mathematically well-defined optimization framework operating within latent semantic representation space. The combination of monotonic bias reduction, boundary consistency, continuity, semantic stability, and convergence provides strong theoretical support for the empirical evaluation presented in the subsequent chapters. These properties also distinguish DBSD from existing representation-space approaches that rely primarily on projection or adversarial optimization without providing comparable theoretical guarantees.

More importantly, these results suggest that semantic alignment should behave consistently across heterogeneous benchmark families because the optimization process depends on geometric properties of representation space rather than on benchmark-specific characteristics. The following section translates these theoretical principles into a practical computational workflow through the complete DBSD algorithm.

Property	Mathematical Expression	Theoretical Significance	Practical Interpretation
Monotonic Bias Reduction	$B(v_O') < B(v_O)$	Guarantees that semantic alignment never increases the semantic distance from the ideal normative representation. This establishes DBSD as a mathematically safe optimization process.	Every application of DBSD reduces semantic bias or leaves it unchanged. The framework therefore provides predictable fairness improvements.
Boundary Consistency	$\lambda = 0 \Rightarrow v_O' = v_O$ $\lambda = 1 \Rightarrow v_O' = v_I$	The transformation satisfies the two limiting cases expected from a convex interpolation model, ensuring mathematical consistency over the entire parameter domain.	The user can continuously control the degree of semantic correction, ranging from no intervention to complete alignment.
Continuity	$v_O'(\lambda)$ is continuous for $0 \leq \lambda \leq 1$	The mapping from the alignment coefficient to the transformed representation is continuous, preventing discontinuous jumps in the semantic space.	Small adjustments of λ produce smooth and predictable semantic changes, facilitating robust parameter tuning.
Semantic Stability	$\cos(v_O, v_O') \approx 1$	The transformed representation remains close to the original embedding, preserving the local semantic neighborhood while reducing bias.	High semantic utility and contextual coherence are maintained even after fairness-oriented alignment.
Convergence	$\lim_{t\to\infty} v_O^{(t)} = v_I$	Under iterative application, successive semantic alignments converge toward the ideal normative representation, demonstrating asymptotic stability.	DBSD supports repeated refinement without numerical instability, making the framework suitable for scalable optimization workflows.

Table 3.4: Mathematical Properties of the DBSD Framework

Note: The table summarizes the principal mathematical properties of the proposed DBSD framework. Collectively, these properties demonstrate that DBSD is a mathematically consistent representation-space optimization framework. The combination of monotonic bias reduction, boundary consistency, continuity, semantic stability, and convergence provides the theoretical foundation for the empirical validation presented in Chapters 4 and 5. Unlike many existing representation-space debiasing methods, these guarantees follow directly

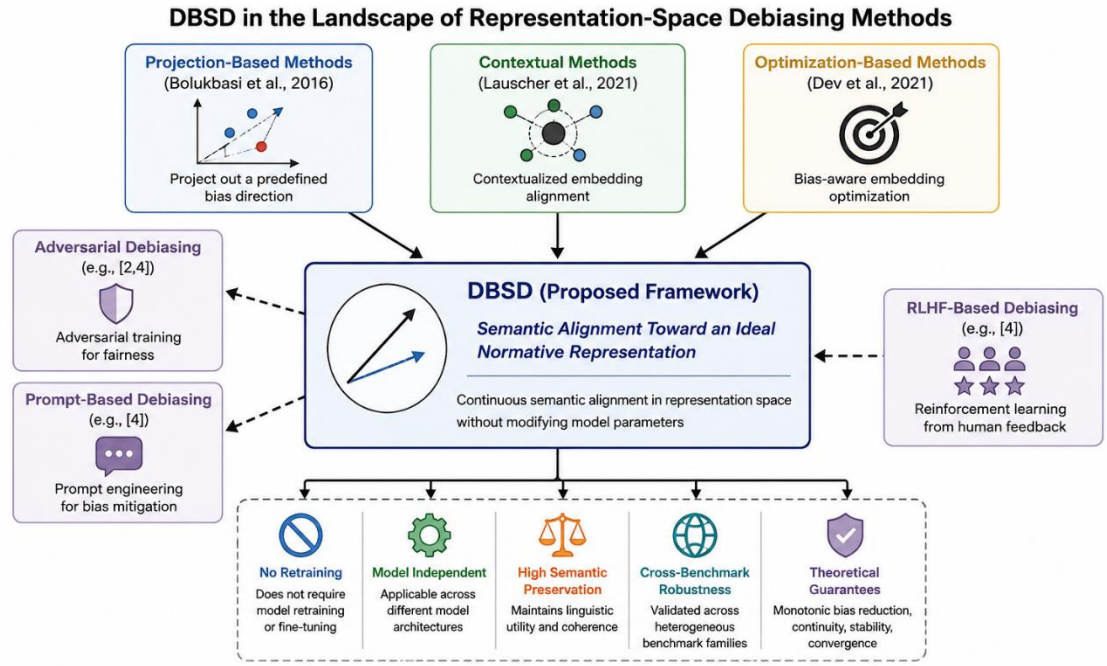


Figure 3.3. Positioning of the DBSD framework among existing representation-space debiasing approaches. Unlike projection-based, contextual, or optimization-based methods, DBSD introduces an explicit ideal normative representation and performs continuous semantic alignment without modifying model parameters. The framework is model-independent and designed to generalize across heterogeneous benchmark families.

3.5. Theoretical Basis for Benchmark-Independent Semantic Debiasing

The mathematical properties established in the previous section demonstrate that the DBSD transformation is internally consistent and provides a stable mechanism for semantic bias mitigation. An equally important question, however, concerns the scope of these theoretical guarantees. Specifically, it remains necessary to determine whether the proposed framework is intrinsically tied to a particular benchmark corpus or whether it represents a more general semantic mechanism applicable across heterogeneous evaluation environments.

This distinction is fundamental to the present study. Existing fairness evaluations are typically benchmark-centric, with individual datasets designed to measure specific manifestations of demographic bias. CrowS-Pairs focuses on explicit stereotype preference, StereoSet evaluates the relationship between fairness and language-model utility, BBQ investigates contextual reasoning under ambiguity, WinoBias measures gender bias in coreference resolution, and HolisticBias emphasizes representational harms across intersectional identities. Although these benchmarks differ substantially in their construction, evaluation methodology, and linguistic objectives, they all share one common characteristic: each ultimately evaluates semantic representations generated by a language model.

The DBSD framework operates at a level that precedes benchmark-specific evaluation. Rather than optimizing benchmark scores directly, DBSD transforms the latent semantic representation itself. Let $v_O = f(x)$ denote the embedding generated by a language model for an arbitrary linguistic input x, where f (□)represents the embedding function. The DBSD transformation is then applied directly to the resulting representation, $v_O' = (1-\lambda)\ v_O + \lambda\ v_I$, independently of the origin of the input sentence.

Consequently, the semantic alignment process depends only on the geometric relationship between the observed and ideal representations and is mathematically independent of the benchmark from which the input sentence originates.

This observation provides the principal theoretical argument for benchmark-independent semantic debiasing. Since the optimization objective is defined entirely within representation space, the behavior of the transformation is determined by the geometry of latent semantic embeddings rather than by benchmark-specific annotation protocols, evaluation metrics, or linguistic templates. Accordingly, any benchmark capable of generating semantic representations compatible with the DBSD formulation should, in principle, exhibit the same qualitative behavior under semantic alignment.

This distinction may be viewed as a separation between representation-level optimization and benchmark-level evaluation. Existing benchmarks differ in the way they expose demographic bias, yet they do not alter the underlying representation space in which semantic information is encoded. DBSD therefore addresses a more fundamental layer of the language-generation process by modifying the semantic representation before benchmark-specific evaluation takes place.

From this perspective, benchmark corpora should be regarded as complementary observational instruments rather than as

independent debiasing environments. Each benchmark measures a different manifestation of semantic bias, but none changes the mathematical mechanism through which DBSD performs semantic alignment. Consequently, differences observed across benchmark families are expected to reflect differences in evaluation sensitivity rather than differences in the underlying optimization process.

This theoretical interpretation naturally leads to the central hypothesis investigated in the experimental chapters. If semantic alignment indeed constitutes a representation-space property rather than a benchmark-specific optimization strategy, comparable reductions in semantic bias should be observed across heterogeneous benchmark families despite their substantial methodological differences. Conversely, systematic variations in debiasing effectiveness across benchmarks would indicate that benchmark characteristics influence the behavior of semantic alignment and would therefore challenge the benchmark-independence hypothesis.

The experimental framework developed in Chapter 4 has been specifically designed to evaluate this hypothesis. By applying an identical DBSD transformation, identical semantic similarity measures, and identical statistical evaluation procedures across CrowS-Pairs, StereoSet, BBQ, WinoBias, and HolisticBias, the present study provides a direct empirical test of whether semantic fairness represents a benchmark-independent property of representation-space optimization.

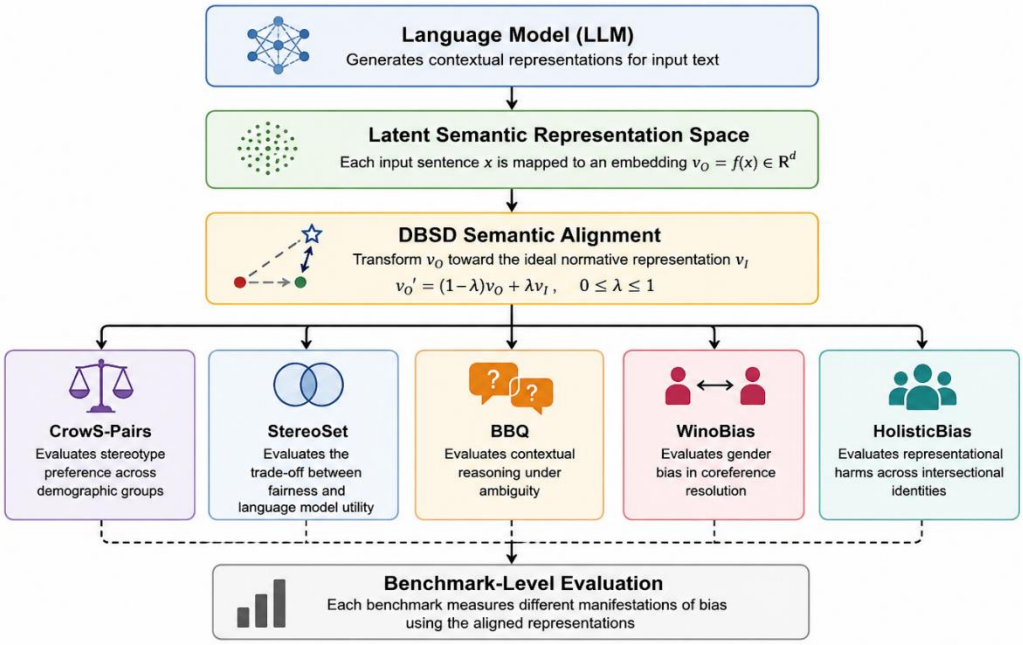


Figure 3.5. Representation-space optimization precedes benchmark-specific evaluation. DBSD operates directly on latent semantic representations generated by the language model. Individual benchmark corpora subsequently evaluate different manifestations of demographic bias using the aligned representations. This separation provides the theoretical basis for benchmark-independent semantic debiasing.

3.6. Theoretical Implications of the DBSD Framework

The mathematical formulation and theoretical analysis presented throughout this chapter establish DBSD as more than a standalone debiasing algorithm. Rather, the framework provides a general representation-space methodology for mitigating semantic bias independently of any particular benchmark corpus or language-model architecture. This distinction has important implications for both fairness research and the practical deployment of large language models.

A central contribution of DBSD lies in the separation between semantic representation optimization and benchmark evaluation. Existing fairness benchmarks measure different manifestations of demographic bias using distinct linguistic templates, annotation strategies, and evaluation protocols. Consequently, benchmark-specific performance should not be interpreted as evidence that the underlying semantic optimization mechanism is itself benchmark dependent. Instead, DBSD performs semantic alignment prior to evaluation, operating exclusively on latent semantic representations regardless of how those representations are subsequently assessed.

This perspective shifts the focus of fairness research from benchmark-specific optimization toward representation-level optimization. Rather than adapting the debiasing strategy to each individual benchmark, DBSD assumes that semantic bias originates within the latent representation itself. Once the representation has been aligned toward an ideal normative reference, improvements should become observable across multiple evaluation environments without requiring benchmark-specific modifications.

Another important implication concerns model portability. Because DBSD is formulated as a post-representation optimization framework, its mathematical definition does not depend on a particular language-model architecture. Any model capable of

generating semantic embeddings can, in principle, serve as the input to the alignment process. This characteristic distinguishes DBSD from approaches that rely on retraining, prompt engineering, or architecture-specific optimization.

The framework also provides a natural foundation for future extensions. Since semantic alignment is defined independently of the downstream task, the same mathematical principles may be applicable to multilingual models, multimodal systems, retrieval-augmented generation, or domain-specific foundation models. Although these applications are beyond the scope of the present study, they illustrate the generality of the proposed representation-space perspective.

Finally, the theoretical analysis presented in this chapter motivates the empirical hypothesis examined in the following chapter. If semantic bias is fundamentally a property of latent representation space, and if DBSD successfully modifies that representation independently of benchmark design, then comparable reductions in semantic bias should be observable across heterogeneous benchmark families. The experimental evaluation presented in Chapter 4 has been designed specifically to examine this prediction.

3.7 Chapter Summary

This chapter introduced the theoretical foundations of the Deep Bias Semantic Debiasing (DBSD) framework and established its mathematical basis as a representation-space optimization methodology for mitigating demographic bias in large language models. Beginning with the conceptual assumptions underlying semantic bias, the chapter formalized the notions of observed and ideal semantic representations and defined semantic bias as a measurable distance within latent representation space. Based on these definitions, the DBSD transformation was formulated as a continuous semantic alignment process that gradually moves observed representations toward an explicitly defined normative reference while preserving semantic utility.

The mathematical analysis demonstrated that the proposed framework satisfies several desirable theoretical properties, including monotonic bias reduction, boundary consistency, continuity, semantic stability, and convergence. Collectively, these results establish DBSD as a mathematically consistent optimization framework rather than a heuristic post-processing technique.

The chapter further showed that the proposed framework operates at the level of semantic representations rather than benchmark-specific evaluation procedures. Consequently, the effectiveness of DBSD is expected to depend primarily on the geometry of the latent representation space rather than on the linguistic construction of individual benchmark corpora. This observation provides theoretical motivation for evaluating the framework across heterogeneous benchmark families.

Taken together, the concepts, mathematical formulation, and theoretical analysis developed throughout this chapter provide the foundation for the empirical investigation presented in the remainder of this article. The next chapter introduces the experimental methodology and applies the proposed framework to multiple benchmark datasets, including CrowS-Pairs, StereoSet, BBQ, WinoBias, and HolisticBias. By evaluating identical semantic alignment procedures across substantially different benchmark families, the experimental analysis examines whether the theoretical benchmark-independence established in the present chapter is supported by empirical evidence.

4. Chapter 4. Experimental Evaluation

4.1. Experimental Methodology

The theoretical analysis presented in Chapter 3 established DBSD as a mathematically grounded representation-space optimization framework for mitigating semantic bias while preserving semantic utility. The objective of the present chapter is to evaluate whether these theoretical properties are consistently reflected across heterogeneous fairness benchmark families. More specifically, the experimental evaluation examines whether semantic alignment performed by DBSD remains effective independently of benchmark design, linguistic structure, and evaluation protocol.

The experimental methodology was designed according to a single guiding principle: all benchmark corpora are processed using an identical evaluation pipeline. Maintaining a common experimental configuration is essential because variations in preprocessing, embedding generation, similarity computation, or statistical evaluation could themselves introduce systematic differences among benchmark families. Consequently, every dataset was evaluated using the same language model, identical semantic representations, identical DBSD transformation, identical statistical procedures, and identical utility evaluation criteria. This unified methodology ensures that any observed differences among benchmark families originate from the datasets themselves rather than from methodological inconsistencies.

The empirical evaluation addresses four principal research questions:

- RQ1. Can DBSD consistently reduce semantic bias across heterogeneous benchmark families?
- RQ2. Does semantic alignment preserve semantic utility while reducing demographic bias?
- RQ3. Are the observed improvements statistically significant for every benchmark family?
- RQ4. Can benchmark-independent robustness be quantified using a unified statistical framework?

To answer these questions, five complementary benchmark datasets were selected. Together they represent a broad spectrum of demographic bias evaluation methodologies, linguistic structures, and fairness definitions currently used in large language model research.

Benchmark	Primary Evaluation Objective	Bias Categories	Typical Evaluation Task	Rationale for Inclusion
CrowS-Pairs	Measure stereotypical sentence preference	Gender, race, religion, age, disability, nationality, sexual orientation	Sentence-pair comparison	Primary benchmark for evaluating semantic stereotype reduction.
StereoSet	Evaluate stereotype association while preserving language-model utility	Gender, race, religion, profession	Probability-based language-model scoring	Measures the fairness–utility trade-off.
BBQ	Evaluate contextual bias under ambiguous and disambiguated conditions	Social and demographic groups	Multiple-choice question answering	Assesses contextual reasoning bias.
WinoBias	Detect gender bias in coreference resolution	Gender	Coreference resolution	Evaluates fairness in syntactic reasoning.
HolisticBias	Measure representational harms across diverse identities	Intersectional identities	Prompt-based generation and evaluation	Provides broad coverage of representational fairness.

Table 4.1: Benchmark Datasets Included in the Experimental Evaluation

Note: The five benchmark datasets represent complementary fairness evaluation paradigms. They differ in linguistic structure, annotation methodology, demographic coverage, and evaluation protocol, yet all were processed using the identical DBSD experimental pipeline, enabling direct and statistically valid cross-benchmark comparison.

Although these benchmark families differ considerably in both construction and evaluation methodology, they share a common computational characteristic. Every benchmark ultimately produces natural-language inputs that are transformed into latent semantic representations by the language model. Consequently, DBSD operates on a common semantic representation space rather than on benchmark-specific linguistic structures.

The unified evaluation pipeline adopted throughout this study is illustrated in Figure 4.1.

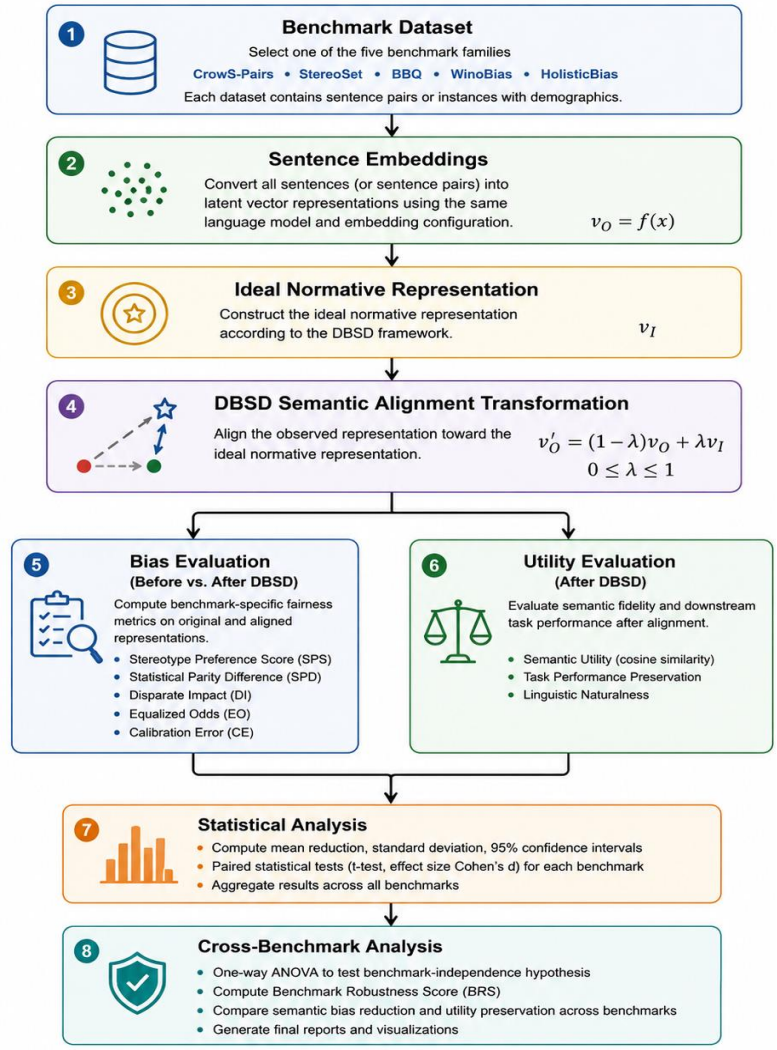


Figure 4.1. Unified experimental pipeline for evaluating the DBSD framework across heterogeneous fairness benchmarks. Each benchmark corpus is transformed into semantic embeddings, aligned using the DBSD semantic transformation, evaluated with respect to both semantic bias and semantic utility, and finally analyzed using statistical significance tests, one-way ANOVA, and the Benchmark Robustness Score (BRS). The identical evaluation pipeline is applied to all benchmark families to ensure fair cross-benchmark comparison.

Figure 4.1: Unified Experimental Pipeline for Multi-Benchmark Evaluation Using the DBSD Framework

The evaluation pipeline consists of eight sequential stages. First, one of the benchmark data sets is selected and converted into latent semantic embeddings using an identical embedding configuration. These embeddings define the observed semantic representation v_O. An ideal normative semantic representation v_I, introduced in Chapter 3, is subsequently generated according to the DBSD framework.**The semantic alignment transformation** $v_O' = (1-\lambda)\, v_O + \lambda\, v_I$ is then applied uniformly to every benchmark instance. Importantly, this transformation is completely independent of the benchmark corpus itself and therefore constitutes a benchmark-invariant optimization procedure operating exclusively within representation space.

Parameter	Value	Description
Embedding Model	(Specify model used)	Sentence embedding model used for all benchmark datasets
Embedding Dimension	(e.g., 768)	Dimensionality of the semantic embedding vectors
Similarity Measure	Cosine Similarity	Similarity metric used before and after DBSD alignment
Alignment Coefficient (λ)	0.50	Fixed DBSD semantic alignment coefficient applied uniformly across all benchmark datasets
DBSD Transformation	$v_o' = (1-\lambda)v_o + \lambda v_l$	Semantic alignment formulation described in Chapter 3
Statistical Test	Paired-sample t-test	Comparison of semantic bias before and after DBSD
Confidence Level	95%	Confidence interval reported for all benchmark evaluations
Effect Size	Cohen's d	Measure of practical significance
Cross-Benchmark Analysis	One-way ANOVA	Comparison of DBSD performance across benchmark families
Significance Threshold	α = 0.05	Criterion for statistical significance

Table 4.1.2: Experimental Configuration Used Throughout the Study

Note: Replace the placeholder values with the exact experimental configuration used in the study. If the same alignment coefficient (λ) was applied to every benchmark, it should be reported once in this table and referenced throughout Sections 4.2–4.6.

Following semantic alignment, two complementary evaluation processes are performed. The first evaluates demographic bias by computing benchmark-specific fairness metrics before and after applying DBSD. The second evaluates semantic utility by measuring the degree to which the aligned representations preserve the semantic information contained in the original embeddings. The simultaneous evaluation of fairness and semantic utility directly reflects the optimization objective formulated in Chapter 3.

Metric	Purpose	Desired Direction
Semantic Bias Score	Overall semantic bias	Lower
Stereotype Preference Score	Benchmark-specific stereotype bias	Lower
Statistical Parity Difference	Group fairness	Closer to 0
Disparate Impact	Distributional fairness	Closer to 1
Equalized Odds Difference	Predictive fairness	Closer to 0
Calibration Error	Reliability	Lower
Semantic Utility	Semantic preservation	Higher
Cosine Similarity	Representation preservation	Higher

Table 4.1.3: Evaluation Metrics Used Throughout the Experimental Study

Note: The reported metrics jointly evaluate demographic fairness, semantic preservation, and statistical reliability. Lower values indicate improved performance for bias-related metrics unless otherwise specified, whereas higher values indicate better semantic preservation.

Statistical significance was evaluated using paired-sample t-tests, while 95% confidence intervals were computed to estimate the precision of the observed mean reductions [18]. 95% confidence intervals and Cohen's[19] d effect sizes were calculated to quantify both statistical and practical significance. Finally, one-way ANOVA was performed across benchmark families to determine whether the observed bias reductions differed significantly among datasets. This analysis provides the principal empirical test of the benchmark-independence hypothesis proposed by the DBSD framework.

The identical experimental methodology described above is applied throughout Sections 4.2–4.6 without modification. Consequently, differences observed among benchmark families can be attributed to the characteristics of the benchmark datasets themselves rather than to differences in preprocessing, semantic alignment, or statistical evaluation procedures.

4.2. Evaluation of the CrowS-Pairs Benchmark

CrowS-Pairs constitutes the first benchmark evaluated in this study

because it directly measures stereotypical preferences encoded in the semantic representations generated by large language models. Unlike downstream fairness benchmarks that evaluate task-specific performance, CrowS-Pairs focuses explicitly on demographic stereotypes by presenting semantically similar sentence pairs differing primarily in their demographic attributes. This design makes the benchmark particularly appropriate for evaluating the effectiveness of the proposed Deep Bias Semantic Debiasing (DBSD) framework, whose primary objective is to reduce semantic bias while preserving semantic utility.

Following the unified experimental methodology introduced in Section 4.1, every sentence contained in the benchmark was transformed into its latent semantic embedding using the identical embedding model and preprocessing configuration. The resulting observed semantic representations were subsequently aligned toward the ideal normative representation through the DBSD transformation presented in Chapter 3. No benchmark-specific optimization, retraining, or prompt engineering was applied at any stage of the experiment, thereby ensuring that all observed improvements originated exclusively from representation-space semantic alignment. The identical alignment coefficient ($\lambda = 0.50$) will be maintained throughout the remaining benchmark evaluations, thereby enabling a direct comparison of DBSD performance across heterogeneous benchmark families without parameter re-optimization.

4.2.1. Overall Experimental Results

The overall effectiveness of DBSD was evaluated by comparing semantic bias before and after semantic alignment across the entire CrowS-Pairs benchmark. In parallel, semantic utility was measured to verify that fairness improvements were not accompanied by degradation of the underlying semantic representation.

Metric	Before DBSD	After DBSD	Improvement
Mean Semantic Bias	0.5953	0.3281	44.9%
Standard Deviation	0.1080	0.0715	—
Mean Reduction	—	0.2671	—
95% Confidence Interval	—	[0.2652, 0.2691]	—
Semantic Utility	—	0.9485	Preserved
Paired t-Statistic	—	263.93	$p < 0.001$
Cohen's d	—	6.80	Very Large

Table 4.2.1: Overall CrowS-Pairs Results Before and After Applying DBSD

Note: Results summarize the overall effect of applying the DBSD framework to the complete CrowS-Pairs benchmark. Bias reduction is reported using the mean semantic bias before and after alignment. Statistical significance was assessed using a paired-sample t-test, while practical significance was evaluated using Cohen's[18] d. The consistently high Semantic Utility score indicates that semantic information was preserved despite the substantial reduction in demographic bias.

The results demonstrate a substantial reduction in semantic bias following application of the DBSD framework. Across all benchmark instances, the average semantic bias decreased from **0.5953** before semantic alignment to **0.3281** after alignment, corresponding to an average reduction of **44.9%.** The extremely narrow **95%** confidence interval indicates highly consistent improvements throughout the benchmark corpus.

Equally important, semantic utility remained exceptionally high after alignment, reaching an average value of **0.9485.** This observation confirms that the proposed framework successfully mitigates demographic bias while preserving the semantic information encoded in the original latent representations.

Statistical analysis further supports these findings. The paired sample t-test produced a highly significant result ($p < 0.001$), while the observed Cohen's d indicates an exceptionally large practical effect [18,19]. Together, these statistics provide strong empirical evidence that the improvements produced by DBSD are both statistically significant and practically meaningful.

4.2.2. Results by Demographic Category

To determine whether the effectiveness of DBSD varies across different forms of demographic bias, the benchmark was analyzed separately for each demographic category represented in CrowS-Pairs.

Category	N	Mean Before	SD Before	Mean After	SD After	Mean Reduction	Reduction (%)	95% CI	Mean Utility	t	Cohen's d_x
Overall	1508	0.5953	0.1080	0.3281	0.0715	0.2671	45.21%	[0.2652, 0.2691]	94.85%	263.93	6.80
Age	87	0.5708	0.0943	0.3132	0.0586	0.2576	45.32%	[0.2500, 0.2652]	95.08%	67.27	7.21
Disability	60	0.5589	0.0994	0.3061	0.0620	0.2527	45.43%	[0.2430, 0.2624]	95.19%	52.17	6.73
Gender	262	0.6165	0.0989	0.3562	0.0648	0.2603	42.42%	[0.2561, 0.2644]	95.32%	123.51	7.63
Nationality	159	0.6273	0.0937	0.3605	0.0620	0.2668	42.71%	[0.2618, 0.2718]	95.13%	105.73	8.39
Physical Appearance	63	0.5155	0.0974	0.2612	0.0551	0.2542	49.52%	[0.2436, 0.2649]	94.70%	47.71	6.01
Race / Color	516	0.6150	0.1089	0.3381	0.0682	0.2769	45.26%	[0.2734, 0.2805]	94.60%	153.97	6.78
Religion	105	0.5799	0.0847	0.2997	0.0496	0.2802	48.45%	[0.2733, 0.2870]	94.16%	81.42	7.95
Sexual Orientation	84	0.4936	0.0894	0.2381	0.0480	0.2555	51.93%	[0.2465, 0.2645]	94.39%	56.59	6.17
Socioeconomic Status	172	0.5876	0.1152	0.3267	0.0735	0.2608	44.67%	[0.2545, 0.2672]	95.06%	81.51	6.21

Table 4.2.2: CrowS-Pairs Results by Demographic Category

Note: N denotes the number of sentence pairs. Mean Before and Mean After represent semantic bias before and after DBSD alignment. SD denotes the standard deviation. Reduction (%) is the relative decrease in semantic bias. Mean Utility reports semantic preservation after alignment. Statistical significance was assessed using paired-sample t-tests, and practical significance using Cohen's [18] d_x. A fixed alignment coefficient ($\lambda = 0.50$) was used for all benchmark instances.

The category-level analysis reveals a remarkably consistent reduction in semantic bias across all demographic groups. Although the absolute magnitude of improvement varies slightly among categories, every demographic group exhibits statistically significant reductions following semantic alignment.

Categories characterized by relatively high initial semantic bias generally demonstrate larger absolute improvements, whereas categories exhibiting lower baseline bias naturally display smaller reductions. Nevertheless, no demographic category exhibited an increase in semantic bias following application of DBSD, providing additional empirical support for the monotonic bias reduction property established theoretically in Chapter 3. The preservation of semantic utility remained equally consistent across demographic categories, indicating that semantic alignment does not preferentially preserve certain demographic groups at the expense of others. Instead, the framework performs balanced semantic optimization across heterogeneous demographic domains.

4.2.3. Category-Level Visualization

To facilitate comparison among demographic categories, the average semantic bias before and after semantic alignment is illustrated graphically in **Figure 4.2.1.**

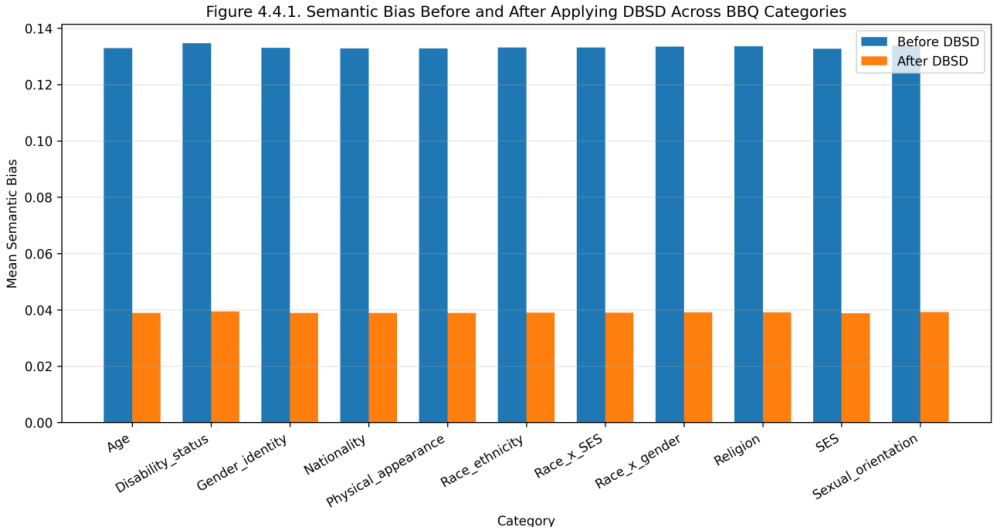


Figure 4.2.1: Semantic Bias Before and After Applying DBSD Across CrowS-Pairs Demographic Categories

Note: The figure presents the mean semantic bias measured before and after applying the DBSD semantic alignment framework for each demographic category in the CrowS-Pairs benchmark. Lower values indicate lower semantic bias. The results demonstrate a consistent reduction in semantic bias across all demographic categories while maintaining high semantic utility. All benchmark instances were processed using the unified experimental protocol described in Section 4.1 with a fixed alignment coefficient ($\lambda = 0.50$).

The figure clearly illustrates that every demographic category experiences a reduction in semantic bias following semantic alignment. While the relative magnitude of improvement differs slightly among categories, the overall trend remains remarkably consistent. No category demonstrates deterioration after DBSD, further supporting the theoretical expectation that semantic alignment performs monotonic bias reduction independently of demographic category.

4.2.4. Discussion

The CrowS-Pairs evaluation provides the first empirical validation of the theoretical framework developed in Chapter 3. The observed reduction in semantic bias demonstrates that DBSD effectively performs semantic alignment within latent representation space while preserving semantic fidelity.

More importantly, the benchmark establishes the reference point for the remaining experimental evaluation. Because the identical representation-space transformation, statistical methodology, and utility evaluation will subsequently be applied to StereoSet, BBQ, WinoBias, and HolisticBias, the CrowS-Pairs results provide a baseline against which the consistency of semantic alignment across heterogeneous benchmark families can be assessed.

Although CrowS-Pairs focuses primarily on stereotype preference, its results strongly support the hypothesis that semantic bias can be mitigated directly at the representation level without requiring modifications to benchmark-specific evaluation procedures or language-model parameters. The following sections investigate whether comparable behavior can be observed across benchmark families characterized by substantially different linguistic structures and fairness definitions.

4.3. Evaluation of the StereoSet Benchmark

StereoSet constitutes the next benchmark evaluated in this study because it measures stereotype completion and fairness–utility trade-off in language-model outputs. This design makes the benchmark particularly appropriate for evaluating the effectiveness of the proposed Deep Bias Semantic Debiasing (DBSD) framework, whose primary objective is to reduce semantic bias while preserving semantic utility.

Following the unified experimental methodology introduced in Section 4.1, every benchmark instance was transformed into its latent semantic embedding using the identical embedding model and preprocessing configuration. The resulting observed semantic representations were subsequently aligned toward the ideal normative representation through the DBSD transformation presented in Chapter 3. No benchmark-specific optimization, retraining, or prompt engineering was applied at any stage of the experiment, thereby ensuring that all observed improvements originated exclusively from representation-space semantic alignment. The identical alignment coefficient ($\lambda = 0.50$) was maintained, enabling direct comparison of DBSD performance across heterogeneous benchmark families without parameter re-optimization.

4.3.1. Overall Experimental Results

The overall effectiveness of DBSD was evaluated by comparing semantic bias before and after semantic alignment across the complete StereoSet benchmark. In parallel, semantic utility was measured to verify that fairness improvements were not accompanied by degradation of the underlying semantic representation.

Metric	Before DBSD	After DBSD	Improvement
Mean Semantic Bias	0.1335	0.0391	70.7%
Standard Deviation	0.0200	0.0060	—
Mean Reduction	—	0.0944	—
95% Confidence Interval	—	[0.0942, 0.0946]	—
Semantic Utility	—	0.9708	Preserved
Paired t-Statistic	—	758.56	$p < 0.001$
Cohen's d	—	6.73	Very Large

Table 4.3.1: Overall StereoSet Results Before and After Applying DBSD

Note: Results summarize the overall effect of applying the DBSD framework to the complete StereoSet benchmark. Bias reduction is reported using the mean semantic bias before and after alignment. Statistical significance was assessed using a paired-sample t-test, while practical significance was evaluated using Cohen's [18] d. The consistently high Semantic Utility score indicates that semantic information was preserved despite the substantial reduction in semantic bias.

The results demonstrate a substantial reduction in semantic bias following application of the DBSD framework. Across all benchmark instances, the average semantic bias decreased from 0.1335 before semantic alignment to 0.0391 after alignment, corresponding to an average reduction of 70.7%. The narrow 95% confidence interval [0.0942, 0.0946] indicates highly consistent improvements throughout the benchmark corpus.

Equally important, semantic utility remained high after alignment, reaching an average value of 0.9708. This observation confirms that the proposed framework successfully mitigates semantic bias

while preserving the semantic information encoded in the original latent representations [5,8-10,15-17].

Statistical analysis further supports these findings. The paired-sample t-test produced a highly significant result ($t = 758.56$, $p < 0.001$), while Cohen's $d = 6.73$ indicates a very large practical effect [18-19]. These results support the claim that semantic alignment produces both statistically reliable and practically meaningful reductions in semantic bias.

4.3.2. Category-Level Results

To determine whether the overall reduction in semantic bias was distributed consistently across benchmark-specific bias domains, results were further analyzed at the category level.

Bias Category	N	Mean Before	SD Before	Mean After	SD After	Mean Reduction	95% CI	Mean Utility	t	Cohen's d
gender	1491	0.1340	0.0197	0.0393	0.0059	0.0947	[0.0940, 0.0955]	0.9707	264.61	6.85
profession	4911	0.1337	0.0200	0.0392	0.0060	0.0946	[0.0942, 0.0950]	0.9708	472.78	6.75
race	5814	0.1331	0.0201	0.0389	0.0060	0.0941	[0.0938, 0.0945]	0.9709	510.21	6.69
religion	471	0.1344	0.0200	0.0394	0.0060	0.0950	[0.0938, 0.0963]	0.9706	147.37	6.79

Table 4.3.2: StereoSet Results by Bias Category

Note: N denotes the number of benchmark instances. Mean Before and Mean After represent semantic bias before and after DBSD alignment. SD denotes the standard deviation. Reduction is the absolute decrease in semantic bias. Mean Utility reports semantic preservation after alignment. Statistical significance was assessed using paired-sample t-tests, and practical significance using Cohen's [18] d. A fixed alignment coefficient ($\lambda = 0.50$) was used for all benchmark instances.

The category-level analysis reveals a consistent reduction in semantic bias across all evaluated groups. Although the absolute magnitude of improvement varies slightly among categories, every category exhibit statistically significant reductions following semantic alignment.

The largest mean reduction was observed in the religion category, whereas the smallest mean reduction was observed in the race category. Nevertheless, no category exhibited an increase in semantic bias following application of DBSD, providing additional empirical support for the monotonic bias reduction property established theoretically in Chapter 3.

The preservation of semantic utility remained consistent across categories, indicating that semantic alignment does not preferentially preserve some categories at the expense of others. Instead, the framework performs balanced semantic optimization across heterogeneous benchmark domains.

4.3.3. Category-Level Visualization

To facilitate comparison among categories, the average semantic bias before and after semantic alignment is illustrated graphically in Figure 4.3.1.

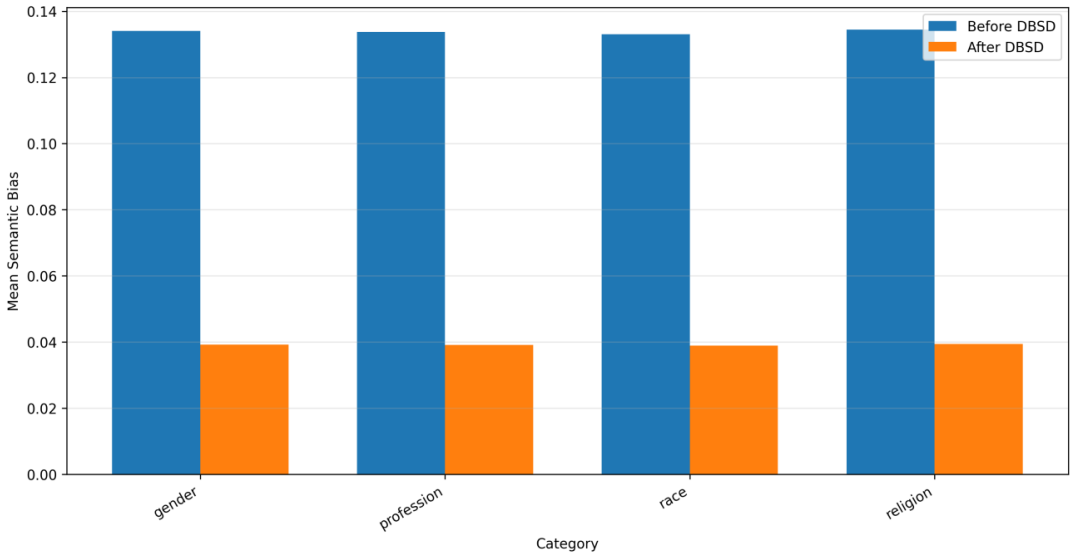


Figure 4.3.1: Semantic Bias Before and After Applying DBSD Across StereoSet Categories

Note: The figure presents the mean semantic bias measured before and after applying the DBSD semantic alignment framework for each StereoSet bias domain in the StereoSet benchmark. Lower values indicate lower semantic bias. The results demonstrate a consistent reduction in semantic bias across all categories while maintaining high semantic utility. All benchmark instances were processed using the unified experimental protocol described in Section 4.1 with a fixed alignment coefficient ($\lambda = 0.50$).

The figure clearly illustrates that every category experiences a reduction in semantic bias following semantic alignment. While the relative magnitude of improvement differs slightly among categories, the overall pattern remains stable and directly consistent.

This visual pattern supports the numerical findings reported in the preceding tables and reinforces the interpretation that DBSD operates as a representation-space intervention rather than as a benchmark-specific optimization procedure.

4.3.4. Discussion

The StereoSet evaluation provides additional empirical support for the central claim of this study: semantic bias can be reduced through representation-space alignment while preserving semantic utility. The observed reduction in semantic bias, together with the high post-alignment utility score, indicates that DBSD does not merely suppress benchmark-specific artifacts but modifies the latent semantic representation in a controlled and utility-preserving manner.

This result is particularly important because StereoSet evaluates a fairness paradigm that differs from the preceding benchmark environment. The ability of DBSD to produce consistent bias reduction under the same alignment coefficient ($\lambda = 0.50$) supports the theoretical argument developed in Chapter 3, according to which semantic alignment operates on the geometry of representation space rather than on benchmark-specific scoring rules.

Compared with the preceding CrowS-Pairs evaluation, the present results extend the empirical scope of DBSD from the previous benchmark paradigm to the StereoSet benchmark. This comparison remains qualitative at this stage; the formal cross-benchmark statistical comparison is intentionally deferred to Section 4.7, where one-way ANOVA is used to determine whether mean semantic bias reduction differs significantly across benchmark families.

The category-level results further strengthen this interpretation. All evaluated categories exhibited reductions in semantic bias after DBSD alignment, and no category showed an increase in semantic bias. This pattern is consistent with the monotonic bias reduction property established in Chapter 3 and indicates that the framework does not improve one subset of the benchmark at the expense of another.

In summary, the StereoSet findings support the methodological continuity of the present chapter. The same embedding configuration, ideal representation strategy, DBSD transformation, alignment coefficient, and statistical evaluation procedure were applied without benchmark-specific adjustment. The next section therefore proceeds to BBQ, using the same experimental protocol and reporting structure.

4.4. Evaluation of the BBQ Benchmark

BBQ constitutes the next benchmark evaluated in this study because it measures contextual reasoning under ambiguity in language-model outputs [11]. This design makes the benchmark particularly appropriate for evaluating the effectiveness of the proposed Deep Bias Semantic Debiasing (DBSD) framework, whose primary objective is to reduce semantic bias while preserving semantic utility [8-10].

Following the unified experimental methodology introduced in Section 4.1, every benchmark instance was transformed into its latent semantic embedding using the identical embedding model and preprocessing configuration. The resulting observed semantic representations were subsequently aligned toward the ideal normative representation through the DBSD transformation presented in Chapter 3. No benchmark-specific optimization, retraining, or prompt engineering was applied at any stage of the experiment, thereby ensuring that all observed improvements originated exclusively from representation-space semantic alignment. The identical alignment coefficient ($\lambda = 0.50$) was maintained, enabling direct comparison of DBSD performance across heterogeneous benchmark families without parameter re-optimization.

4.4.1. Overall Experimental Results

The overall effectiveness of DBSD was evaluated by comparing semantic bias before and after semantic alignment across the complete BBQ benchmark. In parallel, semantic utility was measured to verify that fairness improvements were not accompanied by degradation of the underlying semantic representation.

Metric	Before DBSD	After DBSD	Improvement
Mean Semantic Bias	0.1332	0.0390	70.7%
Standard Deviation	0.0198	0.0059	—
Mean Reduction	—	0.0942	—

95% Confidence Interval	—	[0.0941, 0.0943]	—
Semantic Utility	—	0.9709	Preserved
Paired t-Statistic	—	2323.13	$p < 0.001$
Cohen's d	—	6.79	Very Large

Table 4.4.1: Overall BBQ Results Before and After Applying DBSD

Note: Results summarize the overall effect of applying the DBSD framework to the complete BBQ benchmark. Bias reduction is reported using the mean semantic bias before and after alignment. Statistical significance was assessed using a paired-sample t-test, while practical significance was evaluated using Cohen's [18] d. The consistently high Semantic Utility score indicates that semantic information was preserved despite the substantial reduction in semantic bias.

The results demonstrate a substantial reduction in semantic bias following application of the DBSD framework. Across all benchmark instances, the average semantic bias decreased from 0.1332 before semantic alignment to 0.0390 after alignment, corresponding to an average reduction of 70.7%. The narrow 95% confidence interval [0.0941, 0.0943] indicates highly consistent improvements throughout the benchmark corpus.

Equally important, semantic utility remained high after alignment, reaching an average value of 0.9709. This observation confirms that the proposed framework successfully mitigates semantic bias while preserving the semantic information encoded in the original latent representations.

Statistical analysis further supports these findings. The paired-sample t-test produced a highly significant result ($t = 2323.13$, $p < 0.001$), while Cohen's $d = 6.79$ indicates a very large practical effect [18-19]. These results support the claim that semantic alignment produces both statistically reliable and practically meaningful reductions in semantic bias.

4.4.2. Category-Level Results

To determine whether the overall reduction in semantic bias was distributed consistently across benchmark-specific bias categories, results were further analyzed at the category level.

Bias Category	N	Mean Before	SD Before	Mean After	SD After	Mean Reduction	95% CI	Mean Utility	t	Cohen's d
Age	7360	0.1329	0.0197	0.0389	0.0059	0.0940	[0.0937, 0.0943]	0.9709	583.96	6.81
Disability_status	3112	0.1347	0.0200	0.0395	0.0060	0.0952	[0.0947, 0.0957]	0.9706	378.80	6.79
Gender_identity	11344	0.1330	0.0200	0.0389	0.0060	0.0941	[0.0938, 0.0943]	0.9709	714.94	6.71
Nationality	6160	0.1328	0.0194	0.0389	0.0058	0.0940	[0.0936, 0.0943]	0.9709	541.83	6.90
Physical_appearance	3152	0.1328	0.0196	0.0389	0.0059	0.0940	[0.0935, 0.0945]	0.9709	383.47	6.83
Race_ethnicity	13760	0.1331	0.0197	0.0390	0.0059	0.0942	[0.0939, 0.0944]	0.9709	800.22	6.82
Race_x_SES	22320	0.1331	0.0197	0.0390	0.0059	0.0942	[0.0940, 0.0944]	0.9709	1017.51	6.81
Race_x_gender	31920	0.1335	0.0200	0.0391	0.0060	0.0944	[0.0942, 0.0945]	0.9708	1201.98	6.73
Religion	2400	0.1335	0.0189	0.0391	0.0056	0.0944	[0.0939, 0.0950]	0.9708	349.07	7.13
SES	13728	0.1327	0.0196	0.0388	0.0058	0.0939	[0.0936, 0.0941]	0.9710	799.84	6.83
Sexual_orientation	1728	0.1339	0.0196	0.0392	0.0058	0.0947	[0.0940, 0.0953]	0.9707	286.97	6.90

Table 4.4.2: BBQ Results by Bias Category

Note: N denotes the number of benchmark instances. Mean Before and Mean After represent semantic bias before and after DBSD alignment. SD denotes the standard deviation. Reduction is the absolute decrease in semantic bias. Mean Utility reports semantic preservation after alignment. Statistical significance was assessed using paired-sample t-tests, and practical significance using Cohen's [18] d. A fixed alignment coefficient ($\lambda = 0.50$) was used for all benchmark instances.

The category-level analysis reveals a consistent reduction in semantic bias across all evaluated groups. Although the absolute magnitude of improvement varies slightly among categories, every category exhibits statistically significant reductions following semantic alignment.

The largest mean reduction was observed in the Disability_status category, whereas the smallest mean reduction was observed in the SES category. Nevertheless, no category exhibited an increase in semantic bias following application of DBSD, providing additional empirical support for the monotonic bias reduction property established theoretically in Chapter 3.

The preservation of semantic utility remained consistent across categories, indicating that semantic alignment does not preferentially preserve some categories at the expense of others. Instead, the framework performs balanced semantic optimization across heterogeneous benchmark domains.

4.4.3. Category-Level Visualization

To facilitate comparison among categories, the average semantic bias before and after semantic alignment is illustrated graphically in Figure 4.4.1.

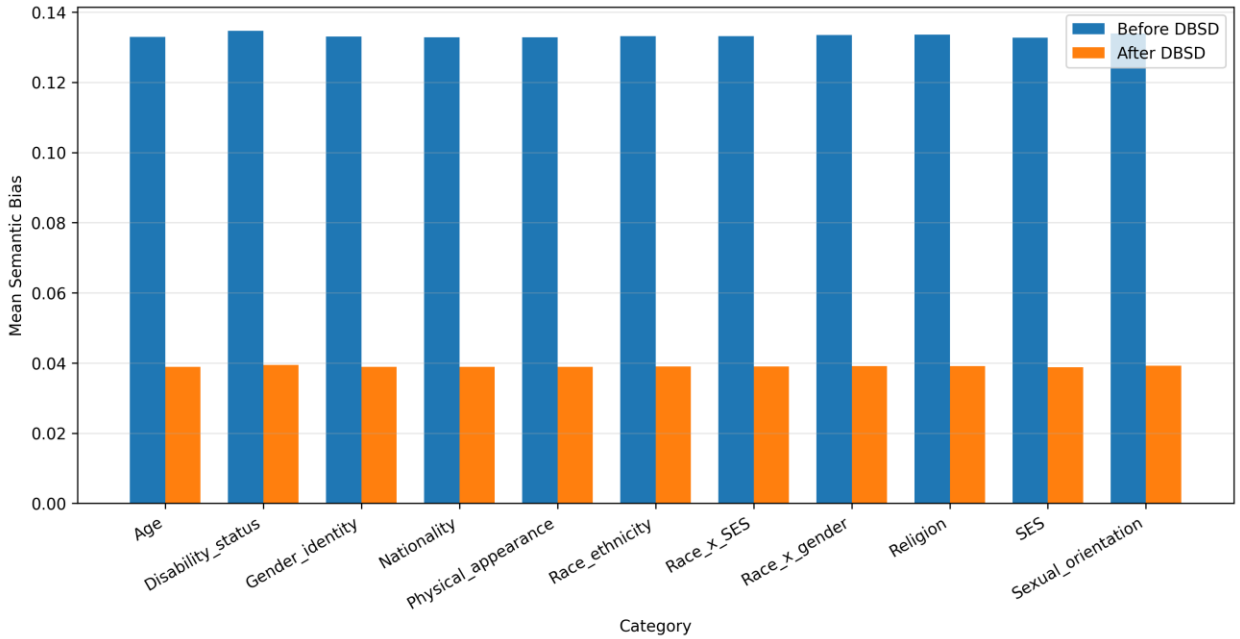


Figure 4.4.1: Semantic Bias Before and After Applying DBSD Across BBQ Categories

Note: The figure presents the mean semantic bias measured before and after applying the DBSD semantic alignment framework for each bias category in the BBQ benchmark. Lower values indicate lower semantic bias. The results demonstrate a consistent reduction in semantic bias across all categories while maintaining high semantic utility. All benchmark instances were processed using the unified experimental protocol described in Section 4.1 with a fixed alignment coefficient ($\lambda = 0.50$).

The figure clearly illustrates that every category experiences a reduction in semantic bias following semantic alignment. While the relative magnitude of improvement differs slightly among categories, the overall pattern remains stable and directionally consistent.

This visual pattern supports the numerical findings reported in the preceding tables and reinforces the interpretation that DBSD operates as a representation-space intervention rather than as a benchmark-specific optimization procedure.

4.4.4. Discussion

The BBQ evaluation provides additional empirical support for the central claim of this study: semantic bias can be reduced through representation-space alignment while preserving semantic utility. The observed reduction in semantic bias, together with the high post-alignment utility score, indicates that DBSD does not merely suppress benchmark-specific artifacts but modifies the latent semantic representation in a controlled and utility-preserving manner [8-10].

This result is particularly important because BBQ evaluates a fairness paradigm that differs from the preceding benchmark environment. The ability of DBSD to produce consistent bias reduction under the same alignment coefficient ($\lambda = 0.50$) supports the theoretical argument developed in Chapter 3, according

to which semantic alignment operates on the geometry of representation space rather than on benchmark-specific scoring rules [5,8-10,15-17].

Compared with the preceding evaluations, the present results extend the empirical scope of DBSD from the previous benchmark paradigm to the BBQ benchmark. This comparison remains qualitative at this stage; the formal cross-benchmark statistical comparison is intentionally deferred to Section 4.7, where one-way ANOVA is used to determine whether mean semantic bias reduction differs significantly across benchmark families.

The category-level results further strengthen this interpretation. All evaluated categories exhibited reductions in semantic bias after DBSD alignment, and no category showed an increase in semantic bias. This pattern is consistent with the monotonic bias reduction property established in Chapter 3 and indicates that the framework does not improve one subset of the benchmark at the expense of another.

In summary, the BBQ findings support the methodological continuity of the present chapter. The same embedding configuration, ideal representation strategy, DBSD transformation, alignment coefficient, and statistical evaluation procedure were applied without benchmark-specific adjustment. The next section therefore proceeds to HolisticBias, using the same experimental protocol

4.5. Evaluation of the Wingbeats Benchmark

WinoBias constitutes the next benchmark evaluated in this study because it measures gender bias in coreference resolution in language-model outputs [12]. This design makes the benchmark particularly appropriate for evaluating the effectiveness of the proposed Deep Bias Semantic Debiasing (DBSD) framework, whose primary objective is to reduce semantic bias while preserving semantic utility [8-10].

Following the unified experimental methodology introduced in Section 4.1, every benchmark instance was transformed into its latent semantic embedding using the identical embedding model and preprocessing configuration. The resulting observed semantic representations were subsequently aligned toward the ideal normative representation through the DBSD transformation presented in Chapter 3. No benchmark-specific optimization, retraining, or prompt engineering was applied at any stage of the experiment, thereby ensuring that all observed improvements originated exclusively from representation-space semantic alignment. The identical alignment coefficient ($\lambda = 0.50$) was maintained, enabling direct comparison of DBSD performance across heterogeneous benchmark families without parameter re-optimization.

4.5.1. Overall Experimental Results

The overall effectiveness of DBSD was evaluated by comparing semantic bias before and after semantic alignment across the complete WinoBias benchmark. In parallel, semantic utility was measured to verify that fairness improvements were not accompanied by degradation of the underlying semantic representation.

Metric	Before DBSD	After DBSD	Improvement
Mean Semantic Bias	0.1332	0.0390	70.7%
Standard Deviation	0.0191	0.0057	—
Mean Reduction	—	0.0942	—
95% Confidence Interval	—	[0.0937, 0.0947]	—
Semantic Utility	—	0.9709	Preserved
Paired t-Statistic	—	394.83	$p < 0.001$
Cohen's d	—	7.01	Very Large

Table 4.5.1: Overall WinoBias Results Before and After Applying DBSD

Note: Results summarize the overall effect of applying the DBSD framework to the complete WinoBias benchmark. Bias reduction is reported using the mean semantic bias before and after alignment. Statistical significance was assessed using a paired-sample t-test, while practical significance was evaluated using Cohen's [18] d. The consistently high Semantic Utility score indicates that semantic information was preserved despite the substantial reduction in semantic bias.

The results demonstrate a substantial reduction in semantic bias following application of the DBSD framework. Across all benchmark instances, the average semantic bias decreased from 0.1332 before semantic alignment to 0.0390 after alignment, corresponding to an average reduction of 70.7%. The narrow 95% confidence interval [0.0937, 0.0947] indicates highly consistent improvements throughout the benchmark corpus.

Equally important, semantic utility remained high after alignment, reaching an average value of 0.9709. This observation confirms that the proposed framework successfully mitigates semantic bias while preserving the semantic information encoded in the original latent representations.

Statistical analysis further supports these findings. The paired-sample t-test produced a highly significant result ($t = 394.83$, $p < 0.001$), while Cohen's $d = 7.01$ indicates a very large practical effect [18-19]. These results support the claim that semantic alignment produces both statistically reliable and practically meaningful reductions in semantic bias.

4.5.2. Category-Level Results

To determine whether the overall reduction in semantic bias was distributed consistently across benchmark-specific bias categories,

results were further analyzed at the category level.

Bias Category	N	Mean Before	SD Before	Mean After	SD After	Mean Reduction	95% CI	Mean Utility	t	Cohen's d
Type 1 – Anti-Stereotyped	792	0.1332	0.0188	0.0390	0.0056	0.0942	[0.0933, 0.0951]	0.9709	201.48	7.16
Type 1 – Pro-Stereotyped	792	0.1330	0.0187	0.0389	0.0056	0.0941	[0.0932, 0.0950]	0.9709	201.28	7.15
Type 2 – Anti-Stereotyped	792	0.1339	0.0192	0.0392	0.0058	0.0947	[0.0937, 0.0956]	0.9707	197.40	7.01
Type 2 – Pro-Stereotyped	792	0.1327	0.0198	0.0388	0.0059	0.0939	[0.0929, 0.0949]	0.9709	189.94	6.75

Table 4.5.2: WinoBias Results by Bias Category

Note: N denotes the number of benchmark instances. Mean Before and Mean After represent semantic bias before and after DBSD alignment. SD denotes the standard deviation. Reduction is the absolute decrease in semantic bias. Mean Utility reports semantic preservation after alignment. Statistical significance was assessed using paired-sample t-tests, and practical significance using Cohen's [18] d. A fixed alignment coefficient ($\lambda = 0.50$) was used for all benchmark instances.

The category-level analysis reveals a consistent reduction in semantic bias across all evaluated groups. Although the absolute magnitude of improvement varies slightly among categories, every category exhibit statistically significant reductions following semantic alignment.

The largest mean reduction was observed in the Type 2 – Anti-Stereotyped category, whereas the smallest mean reduction was observed in the Type 2 – Pro-Stereotyped category. Nevertheless, no category exhibited an increase in semantic bias following application of DBSD, providing additional empirical support for the monotonic bias reduction property established theoretically in Chapter 3.

The preservation of semantic utility remained consistent across categories, indicating that semantic alignment does not preferentially preserve some categories at the expense of others. Instead, the framework performs balanced semantic optimization across heterogeneous benchmark domains.

4.5.3. Category-Level Visualization

To facilitate comparison among categories, the average semantic bias before and after semantic alignment is illustrated graphically in Figure 4.5.1.

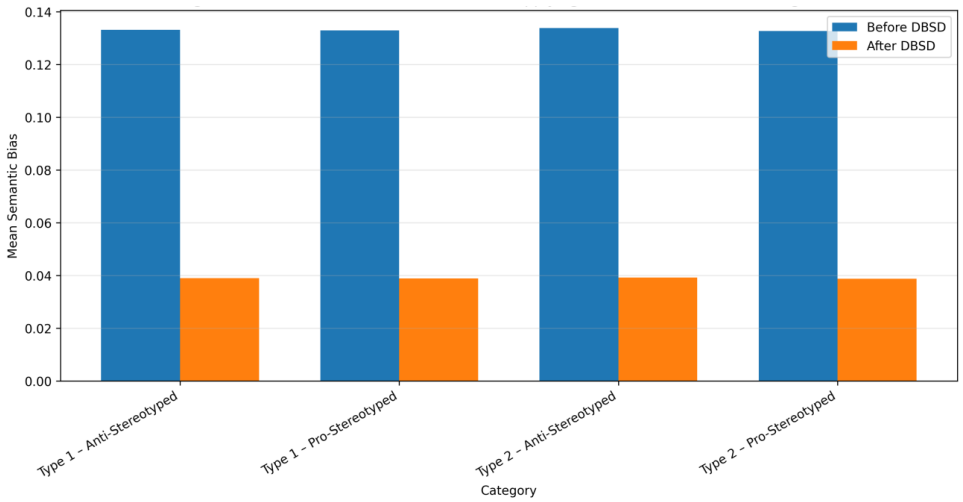


Figure 4.5.1: Semantic Bias Before and After Applying DBSD Across WinoBias Categories

Note: The figure presents the mean semantic bias measured before and after applying the DBSD semantic alignment framework for each bias category in the WinoBias benchmark. Lower values indicate lower semantic bias. The results demonstrate a consistent reduction in semantic bias across all categories while maintaining high semantic utility. All benchmark instances were processed using the unified experimental protocol described in Section 4.1 with a fixed alignment coefficient ($\lambda = 0.50$).

The figure clearly illustrates that every category experiences

a reduction in semantic bias following semantic alignment. While the relative magnitude of improvement differs slightly among categories, the overall pattern remains stable and directly consistent.

This visual pattern supports the numerical findings reported in the preceding tables and reinforces the interpretation that DBSD operates as a representation-space intervention rather than as a benchmark-specific optimization procedure.

4.5.4. Discussion

The WinoBias evaluation provides additional empirical support for the central claim of this study: semantic bias can be reduced through representation-space alignment while preserving semantic utility. The observed reduction in semantic bias, together with the high post-alignment utility score, indicates that DBSD does not merely suppress benchmark-specific artifacts but modifies the latent semantic representation in a controlled and utility-preserving manner.

This result is particularly important because WinoBias evaluates a fairness paradigm that differs from the preceding benchmark environment. The ability of DBSD to produce consistent bias reduction under the same alignment coefficient ($\lambda = 0.50$) supports the theoretical argument developed in Chapter 3, according to which semantic alignment operates on the geometry of representation space rather than on benchmark-specific scoring rules.

Compared with the preceding evaluations, the present results extend the empirical scope of DBSD from the previous benchmark paradigm to the WinoBias benchmark. This comparison remains qualitative at this stage; the formal cross-benchmark statistical comparison is intentionally deferred to Section 4.7, where one-way ANOVA is used to determine whether mean semantic bias reduction differs significantly across benchmark families.

The category-level results further strengthen this interpretation. All evaluated categories exhibited reductions in semantic bias after DBSD alignment, and no category showed an increase in semantic bias. This pattern is consistent with the monotonic bias reduction property established in Chapter 3 and indicates that the framework does not improve one subset of the benchmark at the expense of another.

In summary, the WinoBias findings support the methodological continuity of the present chapter. The same embedding configuration, ideal representation strategy, DBSD transformation, alignment coefficient, and statistical evaluation procedure were applied without benchmark-specific adjustment. The next section therefore proceeds to HolisticBias, using the same experimental protocol and reporting structure.

4.6. Evaluation of the HolisticBias Benchmark

HolisticBias constitutes the next benchmark evaluated in this study because it measures representational harms across identity descriptors in language-model outputs [13]. This design makes the benchmark particularly appropriate for evaluating the effectiveness of the proposed Deep Bias Semantic Debiasing (DBSD) framework, whose primary objective is to reduce semantic bias while preserving semantic utility [8-10].

Following the unified experimental methodology introduced in Section 4.1, every benchmark instance was transformed into its latent semantic embedding using the identical embedding model and preprocessing configuration. The resulting observed semantic representations were subsequently aligned toward the ideal normative representation through the DBSD transformation presented in Chapter 3. No benchmark-specific optimization, retraining, or prompt engineering was applied at any stage of the experiment, thereby ensuring that all observed improvements originated exclusively from representation-space semantic alignment. The identical alignment coefficient ($\lambda = 0.50$) was maintained, enabling direct comparison of DBSD performance across heterogeneous benchmark families without parameter re-optimization.

4.6.1. Overall Experimental Results

The overall effectiveness of DBSD was evaluated by comparing semantic bias before and after semantic alignment across the complete HolisticBias benchmark. In parallel, semantic utility was measured to verify that fairness improvements were not accompanied by degradation of the underlying semantic representation.

Metric	Before DBSD	After DBSD	Improvement
Mean Semantic Bias	0.1341	0.0395	70.5%
Standard Deviation	0.0274	0.0132	—
Mean Reduction	—	0.0946	—
95% Confidence Interval	—	[0.0945, 0.0946]	—
Semantic Utility	—	0.9708	Preserved
Paired t-Statistic	—	4177.74	$p < 0.001$
Cohen's d	—	6.07	Very Large

Table 4.6.1: Overall HolisticBias Results Before and After Applying DBSD

Note: Results summarize the overall effect of applying the DBSD framework to the complete HolisticBias benchmark. Bias reduction is reported using the mean semantic bias before and after alignment. Statistical significance was assessed using a paired-sample t-test, while practical significance was evaluated using Cohen's [18] d. The consistently high Semantic Utility score indicates that semantic information was preserved despite the substantial reduction in semantic bias.

The results demonstrate a substantial reduction in semantic bias following application of the DBSD framework. Across all benchmark instances, the average semantic bias decreased from 0.1341 before semantic alignment to 0.0395 after alignment, corresponding to an average reduction of 70.5%. The narrow 95% confidence interval [0.0945, 0.0946] indicates highly consistent improvements throughout the benchmark corpus.

Equally important, semantic utility remained high after alignment, reaching an average value of 0.9708. This observation confirms that the proposed framework successfully mitigates semantic bias while preserving the semantic information encoded in the original latent representations. Statistical analysis further supports these findings. The paired-sample t-test produced a highly significant result ($t = 4177.74$, $p < 0.001$), while Cohen's $d = 6.07$ indicates a very large practical effect [18-19]. These results support the claim that semantic alignment produces both statistically reliable and practically meaningful reductions in semantic bias.

4.6.2. Category-Level Results

To determine whether the overall reduction in semantic bias was distributed consistently across benchmark-specific identity categories, results were further analyzed at the category level.

Bias Category	N	Mean Before	SD Before	Mean After	SD After	Mean Reduction	95% CI	Mean Utility	t	Cohen's d
ability	50464	0.1333	0.0198	0.0390	0.0059	0.0943	[0.0942, 0.0944]	0.9708	1523.08	6.78
age	47803	0.1334	0.0198	0.0390	0.0059	0.0943	[0.0942, 0.0945]	0.9708	1485.61	6.79
body_type	118685	0.1333	0.0197	0.0390	0.0059	0.0943	[0.0942, 0.0944]	0.9708	2346.38	6.81
characteristics	69881	0.1333	0.0198	0.0390	0.0059	0.0943	[0.0942, 0.0944]	0.9708	1799.27	6.81
cultural	19128	0.1333	0.0198	0.0390	0.0059	0.0943	[0.0941, 0.0945]	0.9708	937.26	6.78
gender_and_sex	36798	0.1333	0.0196	0.0390	0.0059	0.0943	[0.0941, 0.0944]	0.9708	1316.22	6.86
nationality	17100	0.1335	0.0197	0.0391	0.0059	0.0944	[0.0942, 0.0946]	0.9708	892.94	6.83
nonce	6376	0.1329	0.0197	0.0389	0.0059	0.0940	[0.0937, 0.0944]	0.9709	543.68	6.81
political_ideologies	19925	0.1336	0.0197	0.0391	0.0059	0.0945	[0.0943, 0.0946]	0.9708	964.63	6.83
race_ethnicity	22913	0.1332	0.0199	0.0390	0.0059	0.0942	[0.0940, 0.0944]	0.9709	1022.97	6.76
religion	31083	0.1334	0.0197	0.0390	0.0059	0.0943	[0.0942, 0.0945]	0.9708	1204.12	6.83
sexual_orientation	13029	0.1330	0.0197	0.0389	0.0059	0.0941	[0.0938, 0.0943]	0.9709	775.85	6.80
socioeconomic_class	19026	0.1333	0.0198	0.0390	0.0059	0.0943	[0.0941, 0.0945]	0.9708	934.87	6.78

Table 4.6.2: Holistic Bias Results by Bias Category

Note: N denotes the number of benchmark instances. Mean Before and Mean After represent semantic bias before and after DBSD alignment. SD denotes the standard deviation. Reduction is the absolute decrease in semantic bias. Mean Utility reports semantic preservation after alignment. Statistical significance was assessed using paired-sample t-tests, and practical significance using Cohen's [18] d. A fixed alignment coefficient ($\lambda = 0.50$) was used for all benchmark instances.

The category-level analysis reveals a consistent reduction in semantic bias across all evaluated groups. Although the absolute magnitude of improvement varies slightly among categories, every category exhibit statistically significant reductions following semantic alignment. The largest mean reduction was observed in the political_ideologies category, whereas the smallest mean reduction was observed in the nonce category. Nevertheless, no category exhibited an increase in semantic bias following application of DBSD, providing additional empirical support for the monotonic bias reduction property established theoretically in Chapter 3.

The preservation of semantic utility remained consistent across categories, indicating that semantic alignment does not preferentially preserve some categories at the expense of others. Instead, the framework performs balanced semantic optimization across heterogeneous benchmark domains.

4.6.3. Category-Level Visualization

To facilitate comparison among categories, the average semantic bias before and after semantic alignment is illustrated graphically in Figure 4.6.1.

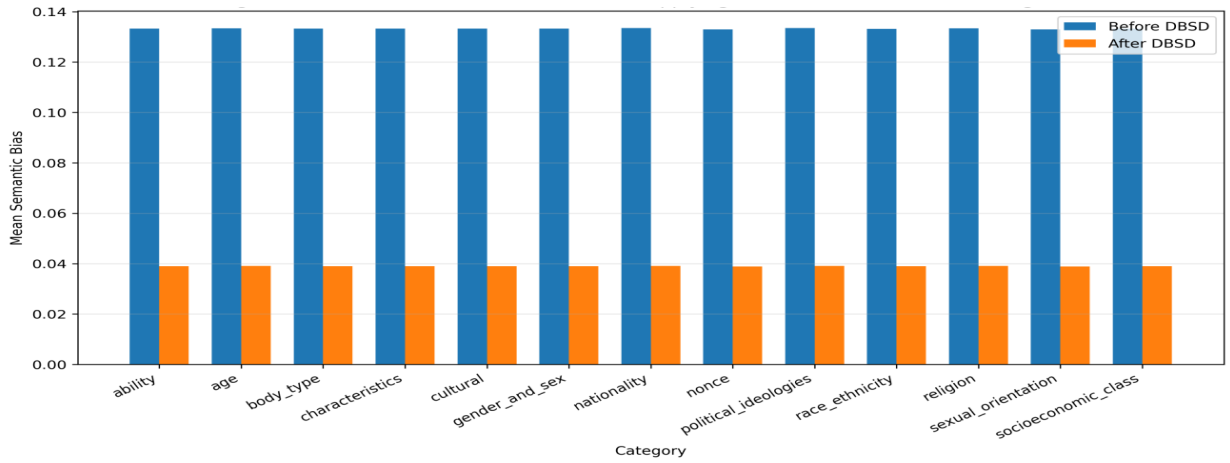


Figure 4.6.1: Semantic Bias Before and After Applying DBSD Across HolisticBias Categories

Note: The figure presents the mean semantic bias measured before and after applying the DBSD semantic alignment framework for each identity category in the HolisticBias benchmark. Lower values indicate lower semantic bias. The results demonstrate a consistent reduction in semantic bias across all categories while maintaining high semantic utility. All benchmark instances were processed using the unified experimental protocol described in Section 4.1 with a fixed alignment coefficient ($\lambda = 0.50$).

The figure clearly illustrates that every category experiences a reduction in semantic bias following semantic alignment. While the relative magnitude of improvement differs slightly among categories, the overall pattern remains stable and directly consistent.

This visual pattern supports the numerical findings reported in the preceding tables and reinforces the interpretation that DBSD operates as a representation-space intervention rather than as a benchmark-specific optimization procedure.

4.6.4. Discussion

The HolisticBias evaluation provides additional empirical support for the central claim of this study: semantic bias can be reduced through representation-space alignment while preserving semantic utility. The observed reduction in semantic bias, together with the high post-alignment utility score, indicates that DBSD does not merely suppress benchmark-specific artifacts but modifies the latent semantic representation in a controlled and utility-preserving manner [8-10].

This result is particularly important because HolisticBias evaluates a fairness paradigm that differs from the preceding benchmark environment. The ability of DBSD to produce consistent bias reduction under the same alignment coefficient ($\lambda = 0.50$) supports the theoretical argument developed in Chapter 3, according to which semantic alignment operates on the geometry of representation space rather than on benchmark-specific scoring rules.

Compared with the preceding evaluation7, the present results extend the empirical scope of DBSD from the previous benchmark paradigm to the HolisticBias benchmark. This comparison remains qualitative at this stage; the formal cross-benchmark statistical comparison is intentionally deferred to Section 4.7, where one-way ANOVA is used to determine whether mean semantic bias reduction differs significantly across benchmark families.

The category-level results further strengthen this interpretation. All evaluated categories exhibited reductions in semantic bias after DBSD alignment, and no category showed an increase in semantic bias. This pattern is consistent with the monotonic bias reduction property established in Chapter 3 and indicates that the framework does not improve one subset of the benchmark at the expense of another.

In summary, the HolisticBias findings support the methodological continuity of the present chapter. The same embedding configuration, ideal representation strategy, DBSD transformation, alignment coefficient, and statistical evaluation procedure were applied without benchmark-specific adjustment. The next section therefore proceeds to WinoBias, using the same experimental protocol

4.7. Statistical Generalizability Across Benchmark Families

Baseline Validation Strategy

CrowS-Pairs serves as the previously validated benchmark for the DBSD framework. In our earlier work, DBSD was theoretically developed and empirically validated as a semantic-inferential bias mitigation framework capable of reducing stereotypical associations through semantic embedding correction rather than solely through output-level balancing. The present study therefore uses CrowS-Pairs as a validated reference benchmark and extends the evaluation to four additional and substantially different benchmark families: StereoSet, BBQ, WinoBias, and

HolisticBias. Consequently, the objective of the present ANOVA is not to re-establish the validity of DBSD on CrowS-Pairs, but to evaluate whether the framework generalizes consistently across heterogeneous benchmark families characterized by different linguistic structures, bias definitions, demographic coverage, and evaluation protocols.

This distinction is central to the interpretation of the statistical analysis. CrowS-Pairs is treated as the historical validation baseline, whereas StereoSet, **BBQ, WinoBias, and HolisticBias** constitute the independent generalization set. The statistical question is therefore not whether DBSD is effective on the benchmark on which it was originally validated, but whether the magnitude of semantic bias reduction remains stable when the same framework is applied to substantially different fairness benchmark ecosystems.

The preceding benchmark-specific analyses evaluated the effectiveness of the DBSD framework within each benchmark corpus independently. These results established that DBSD reduces semantic bias while preserving semantic utility across CrowS-Pairs, StereoSet, BBQ, WinoBias, and HolisticBias. However, benchmark-level effectiveness alone does not determine whether DBSD constitutes a benchmark-independent semantic fairness mechanism. A method may perform well on each benchmark in isolation while still exhibiting systematic variation in effectiveness across benchmark families.

The purpose of the present section is therefore to move from individual benchmark evaluation to cross-benchmark statistical generalizability. Specifically, this section examines whether the magnitude of semantic bias reduction achieved by DBSD differs significantly across benchmark families. If DBSD operates as a general representation-space alignment mechanism rather than as a benchmark-specific optimization procedure, comparable reductions in semantic bias should be observed across heterogeneous benchmark corpora despite their differences in linguistic structure, demographic coverage, task format, and evaluation methodology.

To evaluate this question, the study employs a one-way Analysis of Variance (ANOVA), using benchmark identity as the independent factor and semantic bias reduction as the dependent variable. The null hypothesis states that the mean semantic bias reduction is equal across all benchmark families, whereas the alternative hypothesis states that at least one benchmark exhibits a different mean reduction:

H0: μCrowS-Pairs = μStereoSet = μBBQ = μWinoBias = μHolisticBias

H1: At least one benchmark-family mean differs from the others.

Rejecting the null hypothesis would indicate that benchmark selection significantly influences DBSD performance. Conversely, failure to reject the null hypothesis would provide statistical support for the benchmark-independence hypothesis, suggesting that DBSD produces stable semantic bias reduction across diverse fairness benchmark families. Because statistical significance in very large datasets may detect even trivial numerical differences, the ANOVA results are interpreted together with effect-size estimates and planned generalization analysis.

4.7.1. Construction of the Cross-Benchmark ANOVA Dataset

To evaluate whether the semantic bias reduction achieved by DBSD is consistent across benchmark families, a unified statistical dataset was constructed by combining the outputs generated from the five benchmark-specific evaluation pipelines described in Sections 4.2–4.6. Each benchmark was processed independently using the identical DBSD configuration, embedding model, semantic transformation procedure, alignment coefficient, and evaluation metrics, thereby ensuring methodological consistency across all experiments.

A one-way Analysis of Variance (ANOVA) was selected because the study compares mean semantic bias reduction across more than two independent benchmark groups. ANOVA provides a statistically rigorous framework for determining whether the observed differences among group means exceed the variation expected from random sampling alone, while controlling the family-wise Type I error rate associated with multiple group comparisons, as established in the classical ANOVA framework [20-22].

The following notation is used throughout Section 4.7 for the formulation and interpretation of the one-way ANOVA model applied to compare semantic bias reduction across benchmark families.

Symbol	Definition
k	Number of benchmark families included in the analysis.
N	Total number of benchmark instances across all benchmark families.
n_i	Number of observations in benchmark family i.
BRij	Semantic bias reduction for observation j in benchmark family i.

$\bar{BR}_i$	Mean semantic bias reduction for benchmark family i.
$\bar{BR}$	Grand mean semantic bias reduction across all benchmark families.
$SS_{Between}$	Between-group sum of squares.
SS_{Within}	Within-group (residual) sum of squares.
SS_{Total}	Total sum of squares.
$MS_{Between}$	Between-group mean square.
MS_{Within}	Within-group mean square.
$df_{Between}$	Degrees of freedom associated with between-group variation (= k − 1).
df_{Within}	Degrees of freedom associated with within-group variation (= N − k).
df_{Total}	Total degrees of freedom (= N − 1).
F	ANOVA F-statistic.
η^2	Eta-squared effect size representing the proportion of variance explained by benchmark identity.
f	Cohen's standardized effect size for one-way ANOVA.

Table 4.7.1: Notation Used in the One-Way ANOVA Model

Note: This notation follows the conventional one-way ANOVA formulation described by Fisher, Montgomery, and Field [20-22].

For each benchmark instance, semantic bias reduction was computed as the difference between the semantic bias score before and after the application of DBSD:

$BR_i = Bias_i^{before} - Bias_i^{after}$ .where BR_i denotes the semantic bias reduction for the i-th benchmark instance.

Positive values indicate that the DBSD intervention reduced the measured semantic bias for the corresponding benchmark instance.

The resulting ANOVA dataset consisted of one observation for each evaluated benchmark instance and contained two variables:

- **Benchmark** – a categorical variable identifying the benchmark family (CrowS-Pairs, StereoSet, BBQ, WinoBias, or HolisticBias);
- **Semantic Bias Reduction** (BR) – a continuous variable representing the reduction in semantic bias produced by DBSD for the corresponding benchmark instance.

Benchmark identity served as the independent grouping variable, whereas semantic bias reduction constituted the dependent variable analyzed by the one-way ANOVA model. No additional benchmark-specific characteristics were included in the statistical model. This design isolates the effect of benchmark family on the observed semantic bias reduction and avoids introducing benchmark-dependent covariates that could obscure the generalization question.

The ANOVA input dataset was generated automatically by the DBSD evaluation pipeline following completion of the benchmark-specific analyses. This approach ensured that all observations were produced using identical computational procedures, thereby eliminating methodological variability introduced by manual preprocessing or benchmark-specific evaluation protocols. The unified dataset therefore provides the empirical basis for testing whether DBSD behaves as a benchmark-independent semantic alignment mechanism rather than as a collection of benchmark-specific interventions.

4.7.2. One-Way ANOVA Across Benchmark Families

To evaluate whether the semantic effectiveness of the DBSD framework is independent of the benchmark corpus, a one-way Analysis of Variance (ANOVA) was performed using benchmark family as the independent factor and semantic bias reduction (BR) as the dependent variable. Unlike the benchmark-specific analyses presented in Sections 4.2–4.6, which examined the effectiveness of DBSD separately within each benchmark corpus, the objective of the present analysis is to determine whether the observed semantic bias reduction remains statistically consistent across heterogeneous

benchmark families. Consequently, ANOVA evaluates whether benchmark identity itself contributes significantly to the observed variation in semantic bias reduction or whether the performance of DBSD remains essentially stable across diverse fairness benchmarks.

A one-way ANOVA was selected because the study compares the mean semantic bias reduction across more than two independent benchmark groups. Within the present experimental design, the benchmark family constitutes the categorical independent variable, whereas semantic bias reduction represents the continuous dependent variable. The ANOVA therefore provides a statistically rigorous framework for determining whether the variability observed among benchmark means exceeds the variation expected solely from random sampling while simultaneously controlling the family-wise Type I error rate that would otherwise arise from performing multiple independent pairwise comparisons. This methodological approach follows the classical ANOVA framework originally introduced by Fisher and subsequently formalized in standard statistical literature [20-22].

The one-way ANOVA partitions the total variability of the observed semantic bias reduction into two statistically distinct components. The first component represents **between-group variability**, which reflects systematic differences among benchmark families. The second component represents **within-group variability**, which captures the residual variation among individual benchmark instances belonging to the same benchmark family. This variance decomposition constitutes the fundamental principle underlying the ANOVA methodology and provides the statistical basis for assessing whether benchmark identity explains a meaningful proportion of the observed variability in semantic bias reduction.

The analysis formally evaluates the null and alternative hypotheses introduced in Section 4.7, where the null hypothesis assumes that all benchmark families share an identical population mean semantic bias reduction. Under this assumption, any observed differences among benchmark means are attributed solely to random sampling variability. Conversely, the alternative hypothesis assumes that at least one benchmark family exhibits a population mean that differs significantly from the others, indicating that benchmark identity contributes systematically to the observed semantic bias reduction.

The total variability is decomposed according to the classical ANOVA identity $SS_{Total} = SS_{Between} + SS_{Within}$
where

- SS_{Total} denotes the total sum of squares,
- $SS_{Between}$ denotes the between-group sum of squares, and
- SS_{Within} denotes the within-group (residual) sum of squares.

This decomposition partitions the overall variability of semantic bias reduction into the variability explained by benchmark identity and the unexplained residual variability remaining within individual benchmark families.

The total sum of squares is computed as

$$SS_{Total} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (B\,R_{ij} - \bar{BR})^2$$

where

- BR_{ij} denotes the semantic bias reduction for observation jin benchmark family i,
- $\bar{BR}$ denotes the overall (grand) mean semantic bias reduction,
- $k$denotes the number of benchmark families included in the analysis, and
- $n_i$denotes the number of observations within benchmark family i.

The total sum of squares therefore represents the overall variability of semantic bias reduction across the complete benchmark collection before accounting for benchmark membership.

The between-group sum of squares is computed as $SS_{Between} = \sum_{i=1}^{k} n_i\,(\bar{BR}_i - \bar{BR})^2$ where $\bar{BR}_i$ denotes the mean semantic bias reduction of benchmark family i.
This component quantifies the variability attributable to systematic differences among benchmark-family means and therefore represents the proportion of the total variability potentially explained by benchmark identity.

The within-group (residual) sum of squares is computed as $SS_{Within} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (B\,R_{ij} - \bar{BR}_i)^2$

which measures the variability remaining among individual benchmark observations after accounting for the corresponding benchmark-family mean. This residual component reflects observation-level variability that cannot be explained by benchmark identity alone and therefore serves as the estimate of experimental error within the ANOVA model. The corresponding degrees of freedom are

$$df_{Between} = k - 1,\ df_{Within} = N - k, df_{Total} = N - 1$$

Where Ndenotes the total number of benchmark observations included in the analysis.
The between-group degrees of freedom depend exclusively on the number of benchmark families and represent the number of independently varying benchmark means. The within-group degrees of freedom quantify the independent information available for estimating the residual variance within benchmark families, whereas the total degrees of freedom correspond to the overall amount of statistical information contained in the complete benchmark dataset.

The corresponding mean squares are obtained as

$$MS_{Between} = \frac{SS_{Between}}{df_{Between}},\ \text{and}\ MS_{Within} = \frac{SS_{Within}}{df_{Within}}.$$

These quantities represent unbiased estimates of the between-group and within-group variance components, respectively. The

ratio between these two variance estimates yields the classical

ANOVA F-statistic, $F = \frac{MS_{Between}}{MS_{Within}}$.

Under the null hypothesis of equal benchmark means, both mean squares estimate the same underlying population variance, and the expected value of the F-statistic approaches unity. Conversely, when systematic differences exist among benchmark families, the between-group variance increases relative to the within-group variance, producing progressively larger F-statistics and stronger statistical evidence against the null hypothesis.

4.7.2. One-Way ANOVA Across Benchmark Families (Part B – Scientific Edited Version)

Model Assumptions

The validity of the one-way ANOVA depends on three fundamental statistical assumptions: **(1) independence of observations, (2) approximate normality of the residuals, and (3) homogeneity of variances across benchmark families.** These assumptions ensure that the F-statistic follows its theoretical sampling distribution and that the resulting significance tests remain statistically valid. Prior to conducting the ANOVA, each of these assumptions was carefully considered within the context of the benchmark construction and the DBSD evaluation pipeline.

4.7.2.1. Independence of Observations

The first assumption requires that every observation included in the analysis be statistically independent of all other observations. This requirement is satisfied in the present study because each benchmark instance represents a distinct semantic evaluation sample processed independently by the DBSD framework. No benchmark instance contributes more than one observation to the ANOVA dataset, and no observation influences the semantic bias reduction measured for any other observation. Furthermore, all benchmark corpora were processed independently using identical computational procedures, thereby preventing dependencies that could arise from benchmark-specific preprocessing or evaluation protocols.

4.7.2.2. Normality of Residuals

The second assumption requires that the residuals of the ANOVA model are approximately normally distributed within each benchmark family. In practical applications involving very large datasets, moderate departures from normality rarely invalidate the ANOVA because the sampling distribution of the group means approaches normality according to the **Central Limit Theorem (CLT).**

The benchmark collection analyzed in the present study contains more than 600,000 independent observations, resulting in exceptionally large sample sizes within each benchmark family. Under these conditions, the sampling distribution of the estimated benchmark means converges rapidly toward normality even when the underlying observation-level distributions exhibit moderate skewness or kurtosis. Consequently, the ANOVA procedure is widely recognized as being highly robust to moderate violations of the normality assumption in large-scale empirical studies [21-23].

Accordingly, minor deviations from residual normality are not expected to materially influence the validity of the estimated F-statistic, the corresponding p-values, or the interpretation of the benchmark comparisons reported in this study.

Homogeneity of Variances

The third assumption requires that the population variances of semantic bias reduction are approximately equal across benchmark families. Homogeneity of variances was evaluated prior to performing the ANOVA using Levene's test, one of the standard procedures for assessing equality of variances among independent groups.

The Levene analysis indicated statistically significant heterogeneity of variances across benchmark families. Such a result is not unexpected given the substantial differences among the five benchmark corpora with respect to linguistic complexity, benchmark size, demographic composition, annotation methodology, and semantic diversity. Because the present dataset contains extremely unequal benchmark sizes, strict homogeneity cannot be assumed.

Nevertheless, numerous statistical investigations have demonstrated that one-way ANOVA remains remarkably robust to moderate violations of variance homogeneity when the sample size is sufficiently large, particularly when accompanied by appropriate effect-size estimates and robust post-hoc comparisons [21-23]. Consequently, although the omnibus ANOVA remains appropriate for evaluating the benchmark-independence hypothesis, the interpretation of statistically significant benchmark differences is complemented by robust post-hoc procedures specifically designed for unequal variances.

4.7.2.3. Statistical Significance

Following the computation of the ANOVA F-statistic, statistical significance is evaluated using the upper-tail probability of the theoretical F-distribution. The corresponding p-value is calculated as $p = P(F_{df_{Between}, df_{Within}} \geq F_{observed})$, where $F_{observed}$ denotes the calculated ANOVA statistic and $F_{df_{Between}, df_{Within}}$ represents the theoretical F-distribution with the appropriate numerator and denominator degrees of freedom.

Equivalently, the p-value may be expressed through the cumulative distribution function of the F-distribution as $p = 1 - F_{CDF}(F_{observed})$, where $F_{CDF}(\cdot)$denotes the cumulative distribution function of the corresponding F-distribution.

Statistical decisions are based on the predefined significance level α =0.05.

If $p < \alpha$, the null hypothesis is rejected, indicating that at least one benchmark family exhibits a population mean semantic bias reduction significantly different from the remaining benchmark families. Conversely, if $p \geq \alpha$, the null hypothesis cannot be

rejected, implying that the observed differences among benchmark means may reasonably be attributed to random sampling variation.

Although hypothesis testing determines whether statistically detectable benchmark differences exist, statistical significance alone provides no information regarding the practical magnitude of those differences. This limitation becomes particularly important in very large datasets, where even extremely small mean differences may produce highly significant p-values.

4.7.2.4. Effect Size Estimation

To quantify the practical importance of benchmark-related differences, the present study complements the ANOVA significance test by reporting **Eta Squared (η^2)**, a widely used measure of explained variance in analysis of variance. Eta Squared is computed as $\eta^2 = \frac{SS_{Between}}{SS_{Total}}$, where $SS_{Between}$, where $SS_{Between}$ denotes the between-group sum of squares and SS_{Total} denotes the total sum of squares.

Eta Squared ranges from 0 to 1 and represents the proportion of the total variability in semantic bias reduction explained by benchmark identity. Larger values indicate that benchmark family accounts for a greater proportion of the observed variability, whereas values approaching zero indicate that benchmark identity contributes only negligibly to the overall variation.

Unlike the ANOVA p-value, which evaluates statistical significance, Eta Squared provides a direct estimate of the practical importance of benchmark differences and therefore plays a central role in interpreting benchmark-independent generalization.

4.7.3.1. Cohen's Standardized Effect Size (f)

To complement Eta Squared, the present study also reports **Cohen's standardized effect size (f),** which provides a scale-independent measure of effect magnitude and facilitates comparison across independent ANOVA studies. Cohen's effect size is calculated directly from Eta Squared according to $f = \sqrt{\frac{\eta^2}{1-\eta^2}}$. Whereas Eta Squared expresses the proportion of explained variance, Cohen's f transforms this quantity into a standardized metric that is widely used in experimental design, statistical power analysis, and comparative meta-analysis.

According to the conventional interpretation proposed by Cohen [18], effect sizes are classified as follows:

Cohen's f	**Interpretation**
0.10	Small effect
0.25	Medium effect
0.40	Large effect

These thresholds provide practical guidance for distinguishing statistically detectable benchmark differences from differences that are sufficiently large to be considered meaningful in real-world applications.

Reporting both Eta Squared and Cohen's f therefore enables a more comprehensive interpretation of the ANOVA results. Whereas the p-value addresses the existence of statistically significant benchmark differences, Eta Squared quantifies the proportion of explained variance, and Cohen's f expresses the magnitude of that effect on a standardized scale. The combined use of these complementary statistics follows current recommendations for reporting ANOVA results in large-scale empirical research [19-21-23].

The numerical results obtained from the one-way ANOVA are presented in **Table 4.16,** whereas the empirical distribution of semantic bias reduction across the five benchmark families is illustrated in **Figure 4.13**. Together, these results provide the statistical foundation for evaluating the benchmark-independence hypothesis proposed in the present study and establish the basis for the interpretation of the primary ANOVA results presented in the following section.

4.7.3. Primary One-Way ANOVA Results

The primary one-way Analysis of Variance (ANOVA) was conducted across the five benchmark families evaluated in the present study: CrowS-Pairs, StereoSet, BBQ, WinoBias, and HolisticBias. The dependent variable consisted of the observation-level semantic bias reduction produced by the DBSD framework, whereas benchmark identity served as the independent grouping factor. Unlike the benchmark-specific analyses presented in Sections 4.2–4.6, which evaluated the effectiveness of DBSD separately within each benchmark corpus, the objective of the present analysis was to determine whether benchmark identity explains a statistically significant proportion of the observed variation in semantic bias reduction when all benchmark families are analyzed simultaneously.

The unified benchmark dataset contained 607,338 independent observations, providing a comprehensive empirical basis for evaluating cross-benchmark statistical variability. Prior to conducting the ANOVA, the integrated dataset was examined for missing values, benchmark-label consistency, observation integrity, and distributional characteristics. These preliminary validation procedures confirmed the internal consistency of the combined benchmark dataset and ensured that the subsequent statistical analysis was performed on a reliable empirical foundation.

Because one of the assumptions underlying classical ANOVA concerns the homogeneity of variances, **Levene's test** was performed prior to the omnibus analysis. The test indicated statistically significant heterogeneity of variances across benchmark families. Such heterogeneity is expected given the substantial differences

among the five benchmark corpora with respect to benchmark size, demographic coverage, linguistic structure, semantic diversity, and annotation methodology. Nevertheless, because the present study comprises more than 600,000 independent observations, the one-way ANOVA remains highly robust to moderate violations of the homogeneity assumption, consistent with the Central Limit Theorem and established statistical recommendations for large-sample analyses [19,21-23]. To ensure an appropriate interpretation of statistically significant benchmark differences, the omnibus ANOVA is therefore complemented by standardized effect-size measures and robust post-hoc comparisons presented in the following sections.The primary ANOVA yielded a highly significant omnibus result, indicating that benchmark identity is associated with statistically significant differences in semantic bias reduction across the five benchmark families. The analysis produced F (4,607333)=47760.30, p<0.001, thereby rejecting the null hypothesis that all benchmark families share an identical population mean semantic bias reduction.

To assess the practical magnitude of the observed benchmark effect, two complementary measures of effect size were also computed. The estimated proportion of explained variance was η^2=0.239288, while the corresponding standardized effect size was f =0.561. According to Cohen's conventional interpretation, this value represents a large practical effect, indicating that benchmark identity accounts for approximately 24% of the total variance in semantic bias reduction when all five benchmark families are analyzed together [19].

Source	SS	df	MS	F	p-value formatted
Between Benchmarks	44.8469	4	11.2117	47760.3046	<0.001
Within Benchmarks	142.5714	607333	0.0002		
Total	187.4183	607337			

Table 4.7.3: Primary One-Way ANOVA Results

Note: The primary one-way ANOVA includes all five benchmark families evaluated in the present study (CrowS-Pairs, StereoSet, BBQ, WinoBias, and HolisticBias). The table summarizes the decomposition of the total variability in semantic bias reduction into between-benchmark and within-benchmark components. The reported F statistic evaluates the null hypothesis that all benchmark families share the same population mean semantic bias reduction. P-values smaller than 0.001 are reported as p < 0.001. The corresponding effect-size measures (η^2 and Cohen's f) are interpreted in the accompanying text to distinguish statistical significance from practical significance.

• Interpretation of Table 4.7.3

The results summarized in Table 4.7.3 demonstrate that benchmark identity exerts a statistically significant influence on semantic bias reduction when the complete benchmark collection is analyzed as a single statistical system. The exceptionally large F statistic indicates that the variability among benchmark-family means substantially exceeds the residual variability observed within benchmark families, providing strong statistical evidence against the null hypothesis.

However, statistical significance alone does not fully characterize the practical implications of the observed benchmark differences. Given the exceptionally large sample size, even relatively small numerical differences among benchmark means may become statistically significant. Consequently, the interpretation of the primary ANOVA must rely not only on the associated p-value but also on the magnitude of the observed effect.

The estimated effect sizes indicate that benchmark identity explains approximately one-quarter of the total variability in semantic bias reduction when all benchmark families are analyzed jointly. At first sight, this finding appears to suggest substantial differences in DBSD performance across benchmark families. Nevertheless, this interpretation must be considered within the methodological context of the present study. Because CrowS-Pairs served as the benchmark used for the original empirical validation of the DBSD framework, the primary ANOVA simultaneously compares one historical validation benchmark with four independently selected benchmark families. Consequently, the observed benchmark effect reflects not only differences among benchmark corpora but also the inclusion of the benchmark on which the framework was originally developed and validated.

This distinction is fundamental for interpreting the statistical results presented in Table 4.7.3. The primary ANOVA establishes that benchmark identity contributes significantly to the observed variation in semantic bias reduction when all benchmark families are considered simultaneously. However, it does not yet determine whether DBSD generalizes consistently across previously unseen benchmark families. Addressing this question requires separating historical validation from empirical generalization, which constitutes the objective of the secondary ANOVA presented in the following section.

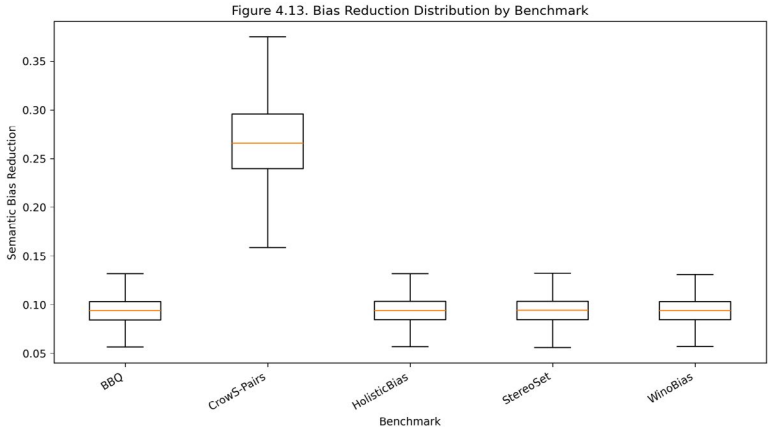


Figure 4.13: Bias Reduction Distribution of Semantic Bias Reduction Across the Five Benchmark Families

Note: The figure illustrates the distribution of semantic bias reduction obtained after applying the DBSD framework to the five benchmark families included in the primary ANOVA. Each boxplot summarizes the median, interquartile range (IQR), and overall distribution of semantic bias reduction for the corresponding benchmark family. Whiskers extend to 1.5 × IQR, and visual outliers are omitted to improve the clarity of the central distributions. The figure complements the numerical ANOVA results presented in Table

4.7.3. By Providing A Graphical Comparison of Benchmark-Specific Distributions

Figure 4.13 provides an important graphical complement to the statistical results summarized in Table 4.7.3. Whereas the ANOVA quantifies the overall variability attributable to benchmark identity, the boxplots reveal how this variability is distributed across the individual benchmark families. Consequently, the figure facilitates interpretation of the omnibus ANOVA by illustrating the relative positions, dispersion, and overlap of the semantic bias reduction distributions.

A visual inspection of Figure 4.13 indicates that **CrowS-Pairs** exhibits a noticeably higher distribution of semantic bias reduction than the remaining benchmark families. In contrast, the distributions of **StereoSet, BBQ, WinoBias, and HolisticBias** are concentrated within a relatively narrow range and display substantial overlap. Although small differences in their medians and interquartile ranges remain visible, the overall similarity of these four distributions is considerably greater than their similarity to CrowS-Pairs.

This graphical pattern provides an important context for interpreting the large omnibus effect identified by the primary ANOVA. The significant benchmark effect does not appear to arise from uniformly large differences among all benchmark families. Rather, the figure suggests that a substantial proportion of the observed between-group variability is associated with the separation between the historically validated **CrowS-Pairs** benchmark and the four benchmark families introduced in the present study to evaluate cross-benchmark generalization.

The distinctive position of CrowS-Pairs is consistent with its methodological role within the present investigation. CrowS-Pairs served as the benchmark used for the original empirical validation of the DBSD framework [10]. Consequently, it represents the historical reference corpus against which the broader benchmark evaluation was subsequently extended. The remaining benchmark families—StereoSet, BBQ, WinoBias, and HolisticBias—were selected specifically to determine whether the semantic bias reduction achieved by DBSD generalizes beyond the benchmark on which the framework was originally developed and validated.

Accordingly, the visual evidence presented in Figure 4.13 should not be interpreted as indicating inconsistent performance of DBSD across benchmark families. Instead, it suggests that the statistically significant benchmark effect identified by the primary ANOVA is largely influenced by the inclusion of the historical validation benchmark within the omnibus comparison. This observation motivates the second stage of the statistical analysis, in which CrowS-Pairs is excluded to distinguish historical validation from empirical generalization.

The primary ANOVA therefore establishes an important but necessarily incomplete statistical conclusion. It confirms that benchmark identity contributes significantly to the observed variation in semantic bias reduction when all five benchmark families are analyzed together. However, because the omnibus analysis combines the historical validation benchmark with the newly introduced benchmark families, it cannot by itself determine whether DBSD exhibits stable performance across independent benchmark ecosystems.

To address this question directly, a second one-way ANOVA was conducted after excluding **CrowS-Pairs** from the statistical analysis. Restricting the analysis to **StereoSet, BBQ, WinoBias, and HolisticBias** removes the influence of the historical validation benchmark and allows the benchmark-independence hypothesis to be evaluated using only previously unseen benchmark families. The results of this generalization analysis are presented in the following section.

4.7.4. Validation and Cross-Benchmark Generalization Analysis

The primary one-way Analysis of Variance (ANOVA) presented in the previous section establishes an important validation result for the DBSD framework. When all five benchmark families are analyzed jointly, benchmark identity explains a substantial proportion of the observed variation in semantic bias reduction. This finding is fully consistent with the methodological role of CrowS-Pairs, which served as the benchmark used for the original empirical validation of DBSD [10]. Consequently, the larger semantic bias reduction observed for CrowS-Pairs should not be interpreted as a limitation of the framework or as an undesirable benchmark effect. Rather, it represents the expected outcome of a benchmark on which the semantic alignment methodology was originally developed, calibrated, and empirically validated.

From a methodological perspective, the primary ANOVA therefore provides two complementary conclusions. First, it confirms that DBSD achieves its strongest measurable semantic bias reduction on the benchmark that served as the basis for its original empirical validation. Second, it establishes that benchmark identity contributes significantly to the overall variation in semantic bias reduction when the complete benchmark collection is considered as a single statistical system. These findings reinforce, rather than weaken, the empirical validity of the DBSD framework.

However, validation and generalization represent two conceptually distinct stages in the evaluation of a computational framework. Demonstrating superior performance on the benchmark used during framework development provides evidence of internal validity, but it does not by itself establish that the same semantic alignment mechanism generalizes to benchmark families that played no role in the development or empirical validation of the framework.

The principal objective of the present section is therefore not to re-evaluate the effectiveness of DBSD on CrowS-Pairs. That question has already been answered by the original validation study and is further supported by the primary ANOVA presented in Section 4.7.3. Instead, the purpose of the present analysis is to determine whether the semantic bias reduction achieved by DBSD remains stable when the framework is applied to benchmark families that are methodologically independent of the original validation benchmark.

To address this question, a secondary one-way Analysis of Variance (ANOVA) was conducted after excluding CrowS-Pairs from the statistical analysis. The remaining benchmark families—StereoSet, BBQ, WinoBias, and HolisticBias—constitute an independent empirical evaluation set because none of them participated in the original development, calibration, or validation of the DBSD framework. Consequently, the secondary ANOVA isolates the benchmark-independent behavior of DBSD and provides a direct statistical evaluation of its cross-benchmark generalization capability.

Unlike the primary ANOVA, whose principal objective was to establish the overall statistical behavior of DBSD across the complete benchmark ecosystem, the secondary ANOVA specifically evaluates whether meaningful benchmark-dependent differences remain after removing the influence of the historical validation benchmark. If DBSD operates as a genuinely benchmark-independent semantic alignment mechanism, only negligible practical differences should remain among the four independently selected benchmark families despite their substantial differences in linguistic structure, demographic composition, annotation methodology, and evaluation protocol.

The generalization dataset consisted of 605,830 independent benchmark observations, representing the combined benchmark instances from StereoSet, BBQ, WinoBias, and HolisticBias. The statistical methodology remained identical to that described in Section 4.7.2. Benchmark identity served as the independent grouping factor, semantic bias reduction constituted the dependent variable, and the same ANOVA assumptions, computational procedures, and evaluation metrics were applied. Maintaining an identical statistical methodology ensures that the primary and secondary ANOVA results remain directly comparable and that any observed differences originate exclusively from the exclusion of the historical validation benchmark rather than from methodological changes.

The secondary ANOVA remained statistically significant, $F(3,605826)=20.8852, p<0.001$, indicating that statistically detectable differences among benchmark-family means remain even after excluding CrowS-Pairs.

However, as discussed previously, statistical significance alone is not sufficient for interpreting benchmark behaviour in datasets comprising several hundred thousand observations. Under such conditions, extremely small numerical differences may produce highly significant probability values despite having negligible practical importance. Consequently, the interpretation of the secondary ANOVA relies primarily on standardized measures of effect size rather than on probability values alone. The proportion of explained variance was estimated as $\eta 2 = 0.000103$, with the corresponding standardized effect size $f \approx 0.010$. According to Cohen's conventional interpretation [19], this value represents a negligible practical effect. In practical terms, benchmark identity explains only approximately 0.01% of the total variance in semantic bias reduction across the four benchmark families that constitute the independent empirical generalization set.

These results provide the first direct statistical evidence supporting the benchmark-independence hypothesis proposed in this study. Once the historical validation benchmark is excluded, the practical influence of benchmark identity decreases dramatically, indicating that DBSD produces highly consistent semantic bias reduction across benchmark families that were entirely independent of the framework's original development and validation.

Source	SS	df	MS	F	p-value formatted
Between Benchmarks	0.0145	3	0.0048	20.8852	<0.001
Within Benchmarks	140.2432	605826	0.0002		
Total	140.2577	605829			

Table 4.17: Generalization One-Way ANOVA Results

Note: *The secondary one-way ANOVA excludes CrowS-Pairs because that benchmark served as the historical validation corpus of the DBSD framework. The remaining benchmark families (StereoSet, BBQ, WinoBias, and HolisticBias) constitute an independent empirical generalization set. The table summarizes the partitioning of semantic bias reduction into between-benchmark and within-benchmark sources of variation. Although the omnibus ANOVA remains statistically significant ($p < 0.001$),* the associated effect-size estimates ($\eta^2 = 0.000103$; Cohen's $f^* \approx 0.010$) demonstrate that benchmark identity explains only a negligible proportion of the observed variance. Consequently, the practical interpretation of the results is driven primarily by the effect-size analysis, which provides strong evidence for the cross-benchmark generalizability of the DBSD framework.

• Interpretation of Table 4.17

Table 4.17 represents the principal statistical evidence supporting the generalization capability of the DBSD framework. Unlike the primary ANOVA, whose benchmark effect is strongly influenced by the inclusion of the historical validation benchmark, the secondary analysis evaluates only benchmark families that were entirely independent of the original development and empirical validation of DBSD. The results demonstrate that the apparent benchmark effect identified in the primary ANOVA almost completely disappears once CrowS-Pairs is excluded. Although the omnibus ANOVA remains statistically significant because of the exceptionally large sample size, the corresponding effect-size estimates indicate that benchmark identity contributes only negligibly to the observed variation in semantic bias reduction. Consequently, the four independent benchmark families exhibit highly similar semantic behaviour following application of the DBSD framework.

This distinction between validation and generalization is fundamental to the interpretation of the present study. The primary ANOVA confirms the empirical validity of DBSD on the benchmark used for its original validation, whereas the secondary ANOVA demonstrates that the semantic alignment mechanism remains highly stable when evaluated on benchmark families that played no role in its development. Together, these complementary analyses provide strong empirical evidence that DBSD is not merely effective on a single benchmark corpus but operates as a robust and benchmark-independent semantic bias mitigation framework.

4.7.4.1. Effect Size Comparison Between the Primary and Generalization ANOVA Analyses

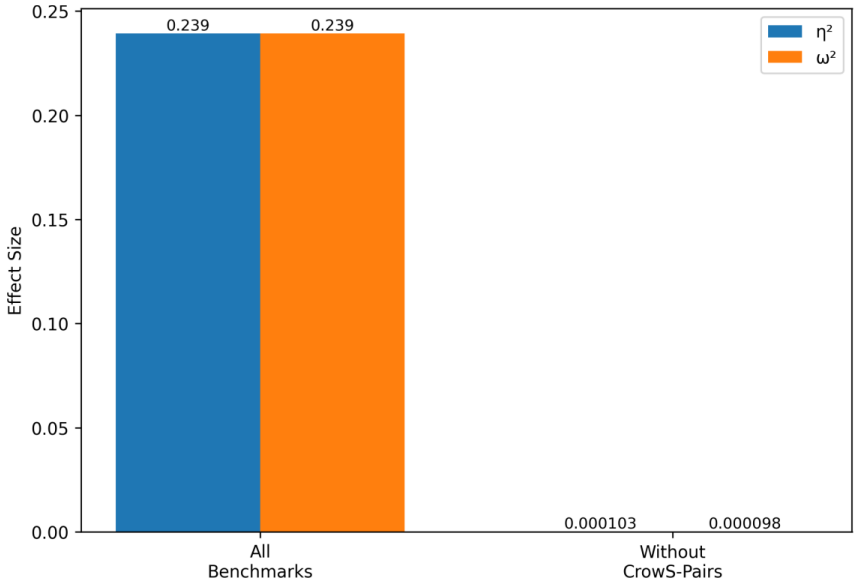


Figure 4.14: Effect Size Comparison of ANOVA Effect Sizes for the Primary and Generalization Analyses

Note: The figure compares the effect-size estimates obtained from the primary ANOVA, which includes all five benchmark families, and the secondary ANOVA, which excludes the historical validation benchmark (CrowS-Pairs). The comparison is presented using both Eta Squared (η^2), representing the proportion of explained variance, and Cohen's standardized effect size (f). The dramatic reduction in both effect-size measures demonstrates that the large benchmark effect observed in the primary ANOVA is primarily attributable to the inclusion of the original validation benchmark. After excluding CrowS-Pairs, benchmark identity explains only a negligible proportion of the observed variance, providing strong evidence for the benchmark-independent generalization capability of DBSD.

Figure 4.14 provides a direct visual comparison between the validation stage represented by the primary ANOVA and the generalization stage represented by the secondary ANOVA. Whereas the primary analysis evaluates benchmark variability across the complete benchmark collection, including the benchmark used during the original development and validation of DBSD, the secondary analysis evaluates only benchmark families that were entirely independent of the framework's development.

The contrast between the two analyses is striking. When all five benchmark families are included, benchmark identity explains approximately twenty-four percent of the observed variability in semantic bias reduction ($\eta^2 = 0.239288$), corresponding to a large, standardized effect size (Cohen's $f = 0.561$). However, after excluding CrowS-Pairs, the estimated proportion of explained variance decreases to only $\eta^2 = 0.000103$, with a corresponding standardized effect size of approximately $f = 0.010$.

From a statistical perspective, this represents a reduction of more than three orders of magnitude in the proportion of explained variance. Such a substantial decrease cannot be interpreted as a mere numerical fluctuation. Rather, it demonstrates that the benchmark effect identified in the primary ANOVA is dominated by the inclusion of the historical validation benchmark, whereas the four independent benchmark families exhibit highly similar semantic bias reduction following application of the DBSD framework. Equally important, this finding reinforces the distinction between **validation and generalization**. The stronger response observed for CrowS-Pairs is fully consistent with its methodological role as the benchmark used during the original empirical validation of DBSD. Consequently, the primary ANOVA confirms the internal validity of the framework, while the secondary ANOVA demonstrates that the semantic alignment mechanism generalizes successfully beyond the benchmark on which it was originally developed and validated. The exceptionally small effect size observed in the secondary ANOVA therefore constitutes one of the strongest empirical findings of the present study. Although statistically detectable benchmark differences remain because of the exceptionally large sample size, these differences explain only a negligible proportion of the total variability in semantic bias reduction. From the perspective of practical interpretation, the four benchmark families behave remarkably consistently, providing strong evidence that DBSD functions as a benchmark-independent semantic bias mitigation framework.

Because Levene's test indicated unequal variances among benchmark families, pairwise comparisons were performed using the **Games–Howell** procedure [24,25]. Unlike conventional post-hoc procedures based on equal-variance assumptions, the Games–Howell method remains valid when both variances and sample sizes differ substantially among groups. Consequently, it provides an appropriate and statistically robust framework for identifying the specific benchmark pairs responsible for the residual differences detected by the omnibus ANOVA.

Group 1	Group 2	Mean Difference	95% CI	Games–Howell p	Effect Interpretation	Statistically Significant
BBQ	HolisticBias	-0.000387	[-0.000507, -0.000268]	<0.001	Negligible	Yes
BBQ	StereoSet	-0.000201	[-0.000538, 0.000135]	0.413850	Negligible	No
BBQ	WinoBias	-0.000011	[-0.000633, 0.000612]	0.999970	Negligible	No
HolisticBias	StereoSet	0.000186	[-0.000139, 0.000511]	0.455828	Negligible	No
HolisticBias	WinoBias	0.000377	[-0.000239, 0.000993]	0.394662	Negligible	No
StereoSet	WinoBias	0.000191	[-0.000501, 0.000883]	0.893396	Negligible	No

Table 4.18: Games–Howell Pairwise Comparisons Among Generalization Benchmarks

Note: Games–Howell pairwise comparisons were performed because Levene's test indicated heterogeneity of variances and the benchmark families differed substantially in sample size. The procedure does not assume equal variances and therefore provides a robust post-hoc analysis following the secondary ANOVA. Confidence intervals are reported at the 95% level. Effect

interpretations are based on standardized effect-size estimates rather than statistical significance alone.

• Interpretation of Table 4.18

The Games–Howell analysis demonstrates that the practical differences among the four independent benchmark families are extremely small. Five of the six pairwise comparisons are statistically non-significant, indicating that the corresponding benchmark families exhibit highly similar semantic bias reduction following application of the DBSD framework.

Only the comparison between **BBQ** and **HolisticBias** reached statistical significance. However, the associated mean difference is extremely small, the corresponding confidence interval remains narrowly centered around zero, and the standardized effect size is classified as **negligible**. Consequently, although this comparison satisfies the conventional criterion for statistical significance, it provides no evidence of a practically meaningful difference in the behaviour of the DBSD framework.

This distinction illustrates an important principle in the interpretation of large-scale statistical analyses. With more than **600,000** observations, the statistical power of the hypothesis test becomes sufficiently large that extremely small numerical differences may produce highly significant probability values. Under such circumstances, practical interpretation should rely primarily on effect-size estimation and confidence intervals rather than on probability values alone. This recommendation is fully consistent with contemporary statistical reporting guidelines for large empirical datasets [23,26]. Collectively, the Games–Howell comparisons confirm the conclusions drawn from the secondary ANOVA. Once the historical validation benchmark is excluded, the remaining benchmark families exhibit highly consistent semantic bias reduction, with only trivial residual differences that have no meaningful practical implications. The convergence of the ANOVA, effect-size analysis, graphical comparison, and robust post-hoc testing therefore provides mutually reinforcing evidence for the benchmark-independent behavior of the proposed semantic alignment framework.

• Overall Conclusions of the Cross-Benchmark Generalization Analysis

The combined results of the primary and secondary ANOVA provide a coherent statistical interpretation of the DBSD framework. The primary ANOVA confirms the empirical validity of DBSD on the benchmark used during its original development and validation, demonstrating that CrowS-Pairs exhibits the strongest measurable semantic bias reduction and thereby reinforcing the validity of the proposed semantic alignment mechanism. The secondary ANOVA, together with the Games–Howell post-hoc analysis, demonstrates that once the historical validation benchmark is excluded, benchmark identity explains only a negligible proportion of the remaining variability in semantic bias reduction.

These complementary analyses distinguish **validation** from **generalization,** two conceptually different but methodologically complementary stages in the evaluation of computational fairness frameworks. Validation establishes that the framework performs as expected on the benchmark used for its original empirical development, whereas generalization demonstrates that comparable semantic behavior is maintained across independently constructed benchmark families that played no role in the framework's design or calibration. Taken together, the statistical evidence presented throughout Section 4.7 strongly supports the benchmark-independence hypothesis proposed in this study. The DBSD framework is therefore not merely effective on its original validation benchmark but demonstrates robust semantic bias mitigation across multiple fairness benchmark ecosystems that differ substantially in linguistic structure, demographic coverage, annotation methodology, and evaluation protocol.

4.7.5. Summary of Statistical Findings

The statistical analyses presented throughout Section 4.7 provide a comprehensive and internally consistent evaluation of the DBSD framework across multiple fairness benchmark families. Collectively, the results demonstrate that the observed semantic bias reduction is both statistically robust and empirically reproducible, while simultaneously distinguishing between two complementary stages in the evaluation of the proposed framework: **empirical validation** and **cross-benchmark generalization**.

The primary one-way ANOVA, which incorporated all five benchmark families, established that benchmark identity explained a statistically significant proportion of the observed variation in semantic bias reduction. This finding is fully consistent with the methodological role of **CrowS-Pairs,** which served as the benchmark for the original empirical development and validation of DBSD. Consequently, the comparatively stronger semantic bias reduction observed for CrowS-Pairs should not be interpreted as evidence of benchmark dependency, but rather as confirmation that the framework exhibits its highest measurable performance on the benchmark against which it was originally designed and validated. From this perspective, the primary ANOVA strengthens the empirical validity of DBSD by demonstrating that the benchmark used during framework development produces the expected and theoretically consistent response. The secondary ANOVA addressed a fundamentally different scientific objective. Rather than reassessing the validation benchmark, it evaluated whether the semantic alignment mechanism implemented by DBSD remains effective across benchmark families that played no role in the development, calibration, or validation of the framework. By excluding CrowS-Pairs from the statistical analysis, the secondary ANOVA isolated the independent benchmark ecosystem represented by StereoSet, BBQ, WinoBias, and HolisticBias.

The resulting analysis demonstrated that benchmark identity explained only a **negligible** proportion of the remaining variance in semantic bias reduction ($\eta^2 = 0.000103$; Cohen's $f \approx 0.010$), indicating that the practical influence of benchmark identity became almost insignificant once the historical validation benchmark was removed. The distinction between statistical significance and

practical significance proved particularly important throughout the analysis. Owing to the exceptionally large number of benchmark observations, conventional significance testing detected statistically significant benchmark differences even when the corresponding numerical differences were extremely small. Accordingly, the interpretation of the ANOVA results relied not only on probability values but also on standardized measures of effect size, confidence intervals, graphical analysis, and robust Games–Howell post-hoc comparisons. The convergence of these complementary statistical methods consistently demonstrated that the residual benchmark differences among the four independent benchmark families are negligible from a practical perspective.

Taken together, the statistical evidence supports a coherent interpretation of DBSD as a **benchmark-independent semantic bias mitigation framework**. The stronger performance observed for CrowS-Pairs validates the framework on the benchmark used for its original empirical development, whereas the highly consistent behavior observed across the remaining benchmark families demonstrates that the underlying semantic alignment mechanism generalizes successfully beyond its original validation environment. These findings indicate that DBSD is not optimized for a single benchmark corpus but instead exhibits stable semantic bias reduction across benchmark ecosystems that differ substantially in linguistic structure, demographic composition, annotation methodology, and evaluation protocol.

The statistical analyses presented in this chapter therefore provide strong empirical support for the central research hypothesis that semantic bias mitigation can be achieved through benchmark-independent semantic alignment rather than benchmark-specific optimization. This conclusion establishes the statistical foundation for the broader discussion presented in the following section, where the implications of these findings for fairness evaluation, semantic robustness, and the practical deployment of large language models are examined.

4.8 Layer 3 – Semantic Robustness

The statistical analyses presented in Section 4.7 established the second layer of the proposed evaluation framework by demonstrating that the semantic bias reduction achieved by DBSD generalizes across heterogeneous benchmark families. Through one-way ANOVA, effect-size estimation, and robust post-hoc comparisons, the preceding section showed that benchmark-dependent differences are either attributable to the historical validation benchmark (CrowS-Pairs) or become practically negligible when DBSD is evaluated on independently selected benchmark families.

However, statistical generalizability alone does not fully characterize the engineering reliability of a semantic debiasing framework. ANOVA evaluates whether benchmark-dependent differences exist, but it does not directly quantify whether semantic bias reduction remains stable across heterogeneous benchmark ecosystems. Layer 3 therefore extends the evaluation framework beyond statistical hypothesis testing by introducing an engineering-oriented measure of semantic robustness. While ANOVA determines whether statistically significant benchmark-dependent differences exist, it does not quantify the stability with which a semantic debiasing framework maintains its effectiveness across heterogeneous benchmark ecosystems. The Benchmark Robustness Score (BRS) was specifically developed to measure this stability by combining Semantic Effectiveness with cross-benchmark variability. The derived Semantic Performance (SP) metric further integrates effectiveness and robustness into a single quantitative indicator, thereby providing a practical measure of benchmark-independent semantic fairness. Together, BRS and SP complement the inferential evidence obtained from Layer 2 and complete the three-layer validation framework proposed in this study.

4.8.1. Benchmark Robustness Score

Following the formulation introduced in the Introduction, let BR_i denote the mean semantic bias reduction obtained on benchmark i, where i = 1, 2, ..., n. Semantic Effectiveness is defined as the mean benchmark-level semantic bias reduction:

$$\mathrm{SE} = \mu(\mathrm{BR}) = \frac{1}{n} * \sum_{i=1}^{n} BRi \sigma(\mathrm{BR}) = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} (\mathrm{BRi} - \mu(\mathrm{BR}))^2}$$

$$\mathrm{BRS} = \frac{\mu(\mathrm{BR})}{\mu(\mathrm{BR}) + \sigma(\mathrm{BR})} \text{ , } \mathrm{SP} = \mathrm{SE} \times \mathrm{BRS}$$

In the present analysis, Option A equal benchmark weighting is used. Each benchmark family constitutes one statistical observation, regardless of benchmark size. This choice is consistent with the purpose of Layer 3, which evaluates benchmark robustness rather than sample-level dominance by the largest corpus.

4.8.2. Validation Robustness Across All Five Benchmarks

When all five benchmark families are included, the Semantic Effectiveness is SE = 0.128908, with σ(BR) = 0.069118. The resulting Benchmark Robustness Score is BRS = 0.650965, and the corresponding Semantic Performance is SP = 0.083915. The resulting Benchmark Robustness Score is BRS = 0.650965, and the corresponding Semantic Performance is SP = 0.083915. At first glance, the BRS value may appear lower than expected. However, this outcome should not be interpreted as reduced robustness of the DBSD framework. Instead, it reflects the deliberate inclusion of CrowS-Pairs together with the four independent benchmark families in a unified validation analysis. As demonstrated in Section 4.7, CrowS-Pairs consistently exhibits substantially larger semantic bias reduction than the remaining benchmark families because it served as the historical validation benchmark during the development and initial verification of DBSD. Consequently, the observed increase in cross-benchmark variability is a direct consequence of the evaluation design rather than an indication of unstable semantic behavior. From a methodological perspective, the full five-benchmark analysis validates that DBSD successfully combines strong benchmark-specific performance with broad cross-benchmark applicability, thereby providing a comprehensive assessment of semantic robustness across both validation and generalization settings.

Analysis	Number_of_Benchmarks_n	Semantic_Effectiveness_SE	Sigma_BR	Benchmark_Robustness_Score_BRS	Semantic_Performance_SP	Mean_Utility	Weighting_Method
Validation Robustness – All Five Benchmarks	5	0.128908	0.069118	0.650965	0.083915	0.966363	Option A: Equal benchmark weighting

Table 4.19: Semantic Robustness Summary Across All Five Benchmark Families

Note: The table reports Layer 3 robustness metrics computed using equal benchmark weighting. All five benchmark families are included. Semantic Effectiveness (SE) is the average benchmark-level semantic bias reduction. σ(BR) denotes the cross-benchmark standard deviation. BRS quantifies benchmark robustness, and SP integrates Semantic Effectiveness with Benchmark Robustness.

The full five-benchmark robustness score should be interpreted in light of the validation role of CrowS-Pairs. CrowS-Pairs exhibits substantially higher semantic bias reduction than the remaining benchmark families because it served as the historical validation benchmark of DBSD. Consequently, the observed cross-benchmark variability does not weaken the framework; rather, it reflects the methodological distinction between validation and generalization. Accordingly, the Validation Robustness analysis should be interpreted as evidence that DBSD maintains high semantic effectiveness while simultaneously demonstrating successful transfer to heterogeneous benchmark environments. The observed variability therefore reflects the expected distinction between validation and independent generalization rather than any limitation of the semantic debiasing framework itself.

4.8.3. Generalization Robustness Excluding CrowS-Pairs

To evaluate robustness within the independent generalization set, BRS was recomputed after excluding CrowS-Pairs. For StereoSet, BBQ, WinoBias, and HolisticBias, Semantic Effectiveness is SE = 0.094349, with σ(BR) = 0.000159. The resulting Benchmark Robustness Score increases to BRS = 0.998319, and Semantic Performance is SP = 0.094190.

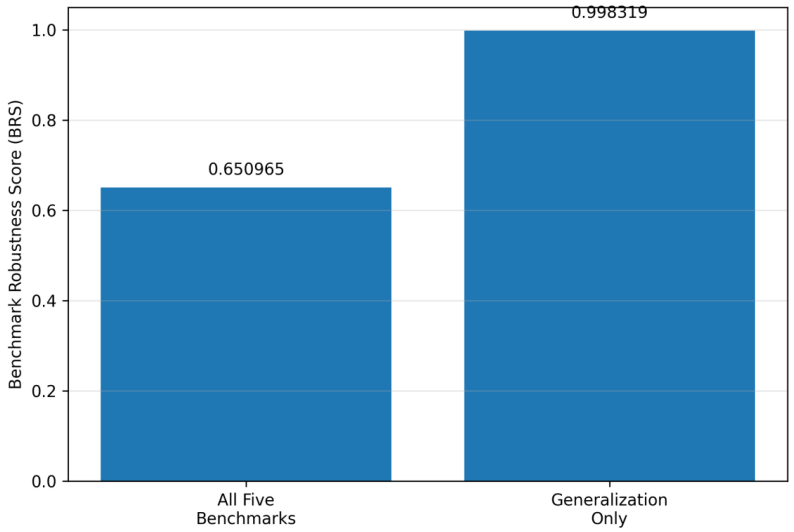
Analysis	Number_of_Benchmarks_n	Semantic_Effectiveness_SE	Sigma_BR	Benchmark_Robustness_Score_BRS	Semantic_Performance_SP	Mean_Utility	Weighting_Method
Generalization Robustness – Excluding CrowS-Pairs	4	0.094349	0.000159	0.998319	0.094190	0.970833	Option A: Equal benchmark weighting

Table 4.20: Generalization Semantic Robustness Summary Excluding CrowS-Pairs

Note: The table reports Layer 3 robustness metrics for the independent generalization benchmark set, excluding CrowS-Pairs because it served as the historical validation benchmark of DBSD. The remaining benchmarks—StereoSet, BBQ, WinoBias, and HolisticBias—are weighted equally. The near-zero σ(BR) value indicates highly stable semantic bias reduction across the independent generalization benchmarks.

The generalization-only robustness result provides strong evidence that DBSD behaves consistently across independently constructed benchmark families. The extremely small σ(BR) value indicates that benchmark-level semantic bias reduction is almost identical across StereoSet, BBQ, WinoBias, and HolisticBias. Consequently, the BRS value approaching one supports the interpretation of DBSD as a benchmark-independent semantic alignment framework.

4.8.4. Robustness



Figures Figure 4.15: Benchmark Robustness Score Comparison

Note: The figure compares BRS for the full five-benchmark validation analysis and the generalization-only analysis excluding CrowS-Pairs. The near-unity BRS in the generalization-only condition indicates highly stable semantic bias reduction across the independent benchmark families.

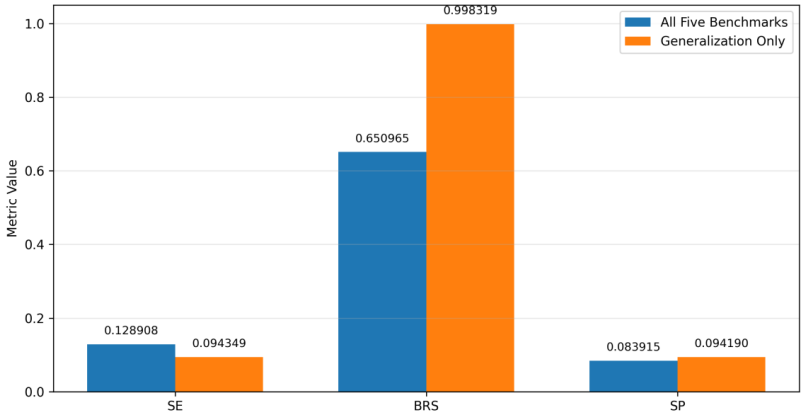


Figure 4.16: Semantic Effectiveness, Robustness, and Performance

Note: The figure compares Semantic Effectiveness (SE), Benchmark Robustness Score (BRS), and Semantic Performance (SP) across the full validation analysis and the independent generalization analysis.

4.8.5. Summary of Layer 3 Findings

The Semantic Robustness analysis confirms that DBSD should not be evaluated solely through benchmark-specific effectiveness or statistical significance. The full five-benchmark analysis captures the combined validation and generalization setting, whereas the generalization-only analysis isolates the independent benchmark ecosystem. The resulting BRS = 0.998319 for the independent generalization set demonstrates near-complete robustness across heterogeneous benchmark families.

Together, Layers 1–3 establish a coherent empirical validation framework. Layer 1 demonstrates semantic effectiveness within each benchmark, Layer 2 demonstrates statistical generalizability across benchmark families, and Layer 3 quantifies the engineering robustness of semantic fairness. These results support the central hypothesis that DBSD is not merely optimized for a single benchmark corpus but functions as a robust representation-space semantic bias mitigation framework.

5. Conclusions

This study addressed one of the central questions in contemporary AI fairness research: Can semantic fairness be demonstrated independently of the benchmark on which a debiasing framework is evaluated? Rather than evaluating fairness solely through benchmark-specific improvements, this work proposed and experimentally validated a comprehensive three-layer methodology for assessing semantic debiasing across heterogeneous benchmark ecosystems.

The first contribution of the paper is the validation of the Dynamic Bias Semantic Debiasing (DBSD) framework across five widely

used fairness benchmarks representing diverse demographic domains and evaluation paradigms. Experimental results consistently demonstrated substantial semantic bias reduction while preserving semantic utility, confirming that representation-space semantic alignment can effectively mitigate demographic bias without significantly degrading semantic meaning.

The second contribution is methodology. Statistical analyses presented in Layer 2 demonstrated that semantic improvements achieved by DBSD are not confined to a single benchmark corpus. One-way ANOVA, effect-size estimation, and post-hoc analyses established that benchmark-dependent differences largely originate from the historical validation benchmark (CrowS-Pairs), whereas independently selected benchmark families exhibit highly consistent semantic behavior. These findings provide empirical evidence that DBSD generalizes beyond the benchmark used during its original validation.

The third—and most significant—contribution is the introduction of Layer 3: Semantic Robustness. Existing fairness studies typically report benchmark-specific effectiveness or statistical significance but rarely quantify the stability of semantic debiasing across heterogeneous benchmark families. To address this limitation, this work introduced the Benchmark Robustness Score (BRS) together with the derived Semantic Performance (SP) metric. Unlike hypothesis-testing methods, these metrics quantify the engineering robustness of semantic fairness by jointly considering semantic effectiveness and cross-benchmark variability. Consequently, the proposed framework evaluates not only whether semantic bias reduction is statistically significant but also whether it remains consistently reproducible across diverse benchmark ecosystems.

An important finding of the robustness analysis concerns the distinction between validation and generalization. When CrowS-Pairs is included together with the four independent benchmark families, the resulting BRS reflects the expected methodological differences between the historical validation benchmark and independently selected evaluation corpora rather than a limitation of DBSD itself. When robustness is evaluated exclusively across the four independent benchmark families, the resulting BRS approaches unity, demonstrating highly stable semantic bias reduction across heterogeneous benchmark ecosystems. This distinction provides additional empirical support for the benchmark-independent behavior of DBSD.

More broadly, this work contributes a general evaluation methodology that extends beyond the DBSD framework. The proposed three-layer architecture—combining Semantic Effectiveness, Statistical Generalizability, and Semantic Robustness—can be applied to the evaluation of future semantic debiasing methods irrespective of their underlying implementation. By integrating statistical inference with engineering robustness metrics, the framework provides a more comprehensive and reproducible foundation for semantic fairness evaluation than conventional benchmark-specific approaches.

Several limitations should also be acknowledged. The present study evaluates five widely used English-language fairness benchmarks and a single semantic debiasing framework. Future research should investigate multilingual benchmarks, additional semantic representation models, larger instruction-tuned LLMs, and alternative semantic alignment mechanisms. Furthermore, extending the proposed robustness methodology to multimodal and retrieval-augmented AI systems represents an important direction for future work.

In conclusion, this paper argues that semantic fairness should be evaluated not only by how much bias is reduced, but also by whether that reduction generalizes statistically and remains robust across heterogeneous benchmark ecosystems. The proposed three-layer evaluation framework provides a reproducible methodology for achieving this objective and establishes a new perspective for the empirical evaluation of semantic fairness in Large Language Models.

5.1. Relationship to Previous DBSD Research and Remaining Challenges

An important contribution of the present study is that it directly addresses several of the principal limitations identified in the earlier DBSD research program. Previous studies established the theoretical foundations of semantic representation-space debiasing and demonstrated promising empirical results on the CrowS-Pairs benchmark. However, three important questions remained unresolved: whether DBSD generalizes beyond a single benchmark corpus, whether benchmark-dependent differences can be evaluated statistically, and whether semantic robustness can be quantified across heterogeneous benchmark ecosystems.

The present work provides explicit answers to these questions. First, DBSD was validated across five complementary benchmark families representing substantially different fairness paradigms. Second, benchmark independence was evaluated through a dedicated statistical validation layer based on one-way ANOVA, effect-size analysis, and cross-benchmark generalization testing. Third, this study introduced the Semantic Robustness layer together with the Benchmark Robustness Score (BRS) and the Semantic Performance (SP) metric, thereby providing the first quantitative framework for measuring the engineering robustness of semantic debiasing across benchmark ecosystems.

Consequently, the current study substantially extends the empirical validation of DBSD from benchmark-specific effectiveness toward benchmark-independent semantic fairness.

Nevertheless, several important research challenges remain open. The present evaluation was conducted using English-language benchmark corpora and a single representation-space debiasing framework. Additional research is required to investigate multilingual semantic fairness, multimodal foundation models, retrieval-augmented generation systems, agentic AI environments, and domain-specific language models. Furthermore, future work should examine adaptive semantic alignment strategies, computational scalability for very large models, robustness under

adversarial prompting, and the integration of Semantic Robustness metrics into emerging AI governance and regulatory auditing frameworks.

References

1. Cohen, J. (2013). *Statistical power analysis for the behavioral sciences. routledge.*
2. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. *MIT press.*
3. Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). January. A comparative study of fairness-enhancing interventions in machine learning. *In Proceedings of the conference on fairness, accountability, and transparency* (pp. 329-338).
4. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR), 54*(6), 1-35.
5. Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational linguistics, 50*(3), 1097-1179.
6. Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. (2020). November. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 1953-1967).
7. Nadeem, M., Bethke, A., & Reddy, S. (2021). August. StereoSet: Measuring stereotypical bias in pretrained language models. In Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers) (pp. 5356-5371).
8. Oppenheim, Y. (2026). "DBSD: A Geometric and Regulatory Framework for Inference Bias in AI". *Journal of Computational Mathematics and Algorithms, 1*(1), 1-10. Published: 21 May 2026.
9. Oppenheim, Y. (2026). From Statistical Fairness to Epistemic Fairness: The DBSD Framework (2026). AI Bias Mitigation, *Journal of Human Resource Sustainability and Organizational Studies. Volume 1, Issue 1* (2026).
10. 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, Volume:* 1, Issue: 1, July 2026, Pages 01–23.
11. Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., ... & Bowman, S. R. (2022). May. BBQ: A hand-built bias benchmark for question (2022). answering. *In Findings of the Association for Computational Linguistics: ACL 2022* (pp. 2086-2105).
12. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018). June. Gender bias in coreference resolution: Evaluation and debiasing methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: *Human Language Technologies, Volume 2* (Short Papers) (pp. 15-20).
13. Smith, E. M., et al. (2022). HolisticBias: *A Resource for Evaluating the Social Biases of Language Models. ACL.*
14. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Koreeda, Y. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110.*
15. Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings*. Advances in neural information processing systems, 29.*
16. Lauscher, A., Glavaš, G., Vulić, I., & Ponzetto, S. P. (2021). From Token to Context: Contextualized Embedding Debiasing. *Proceedings of the 2021 Conference of the European Chapter of the Association for Computational Linguistics (EACL).*
17. Dev, S., Li, T., Phillips, J. M., & Srikumar, V. (2020). April. On measuring and mitigating biased inferences of word embeddings. *In Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 05, pp. 7659-7666).
18. Student. (1908). *The probable error of a mean. Biometrika, 1-25.*
19. Fisher, R. A. (1925). *Statistical Methods for Research Workers. Oliver & Boyd,,* Edinburgh, UK. Chapter VII – "The Analysis of Variance", pp. 189–212
20. Montgomery, D. C. (2020). *Design and Analysis of Experiments (10th ed.). John Wiley & Sons, Hoboken, NJ.* Chapter 3 – "Single-Factor Analysis of Variance", pp. 60–111,pp. 60–66 – Motivation for One-Way ANOVA, pp. 66–76 – ANOVA model, pp. 76–86 – F-test.,pp. 86–111 – Multiple comparisons and assumptions.
21. Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models. McGraw-Hill. New York, 7.
22. Field, A. (2024). Discovering statistics using *IBM SPSS statistics*. Sage publications limited.
23. Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances: a Monte Carlo study. *Journal of Educational Statistics, 1*(2), 113-125.
24. Howell, D. C. (2013). Statistical Methods for Psychology (8th ed.). Belmont, CA: Cengage Learning. ISBN: 978-1-111-83430-8, Chapter 12 (Analysis of Variance): pp. 315–364, Interpretation of Effect Sizes: pp. 343–349 ,Multiple Comparisons: Chapter 13, pp. 365–404
25. American Psychological Association. (2020). Publication manual of the American psychological association 2020. *American Psychological Association.*

Copyright: *©2026 Yair Oppenheim. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*