# Which debiasing approach demonstrates superior robustness in maintaining word embedding coherence during cross-domain semantic

Assignee Research

June 11, 2026

**Abstract**

Contextualized representations (e.g. ELMo, BERT) have become the default pretrained representations for downstream NLP applications. In some settings, this transition has rendered their static embedding predecessors (e.g. Word2Vec, GloVe) obsolete. As a side-effect, we observe that older interpretability methods for static embeddings -while more mature than those available for their dynamic counterparts -are underutilized in studying newer contextualized representations. Consequently, we introduce simple and fully general methods for converting from contextualized representations to static loo

## 1 Introduction

This paper examines: Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. Research question: Which debiasing approach demonstrates superior robustness in maintaining word embedding coherence during cross-domain semantic similarity evaluations?.

## 2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.5/10.

## 3 Results

11 papers retrieved. 9 claims extracted; 9 independently verified. Quality review score: 8.5/10.

## 4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

## 5 Extracted Claims

| Claim | Verified | Confidence |
|---|---|---|
| Contextualized representations (e.g. ELMo, BERT) have become the default pretrained representations for downstream NLP a | ✓ | 0.34 |
| Static embedding predecessors (e.g. Word2Vec, GloVe) are rendered obsolete in some settings due to the transition to con | ✓ | 0.27 |
| Older interpretability methods for static embeddings are more mature than those available for dynamic counterparts. | ✓ | 0.32 |
| The paper introduces simple and fully general methods for converting from contextualized representations to static looku | ✓ | 0.32 |
| The methods are applied to 5 popular pretrained models and 9 sets of pretrained weights. | ✓ | 0.23 |
| Pooling over many contexts significantly improves representational quality under intrinsic evaluation. | ✓ | 0.29 |
| Social biases encoded in pretrained representations with respect to gender, race/ethnicity, and religion are encoded dis | ✓ | 0.39 |
| Models that share the same training data can have different social biases encoded in their representations. | ✓ | 0.20 |
| There are dramatic inconsistencies between social bias estimators for word embeddings. | ✓ | 0.25 |

## References


- https://doi.org/10.1371/journal.pone.0237861
- https://doi.org/10.1007/s10462-024-10896-y
- https://doi.org/10.18653/v1/2020.acl-main.431