Tracing Stereotypes in Pre-trained Transformers: From Biased Neurons to Fairer Models

Gianmario Voria
gvoria@unisa.it
University of Salerno
Fisciano, Italy

Moses Openja
openja.moses@polymtl.ca
Polytechnique Montréal
Montréal, Canada

Foutse Khomh
foutse.khomh@polymtl.ca
Polytechnique Montréal
Montréal, Canada

Gemma Catolino
gcatolino@unisa.it
University of Salerno
Fisciano, Italy

Fabio Palomba
fpalomba@unisa.it
University of Salerno
Fisciano, Italy

Abstract

The advent of transformer-based language models has reshaped how AI systems process and generate text. In software engineering (SE), these models now support diverse activities, accelerating automation and decision-making. Yet, evidence shows that these models can reproduce or amplify social biases, raising fairness concerns. Recent work on neuron editing has shown that internal activations in pre-trained transformers can be traced and modified to alter model behavior. Building on the concept of *knowledge neurons*—neurons that encode factual information—we hypothesize the existence of *biased neurons* that capture stereotypical associations within pre-trained transformers.

To test this hypothesis, we build a dataset of *biased relations*, i.e., triplets encoding stereotypes across nine bias types, and adapt neuron attribution strategies to trace and suppress biased neurons in *BERT* models. We then assess the impact of suppression on SE tasks. Our findings show that biased knowledge is localized within small neuron subsets, and suppressing them substantially reduces bias with minimal performance loss. This demonstrates that bias in transformers can be traced and mitigated at the neuron level, offering an interpretable approach to fairness in SE.

CCS Concepts

• **Software and its engineering → Extra-functional properties**.

Keywords

Fairness; Transformers; Software Engineering

ACM Reference Format:
Gianmario Voria, Moses Openja, Foutse Khomh, Gemma Catolino, and Fabio Palomba. 2018. Tracing Stereotypes in Pre-trained Transformers: From Biased Neurons to Fairer Models. In *Proceedings of IEEE/ACM International Conference on Mining Software Repositories (MSR'26).* ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

MSR'26, Rio de Janeiro, Brazil

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The widespread success of deep learning, and particularly the emergence of transformer architectures [36], has enabled *language models (LMs)*, such as BERT [9] and GPT [23], to become core components of modern AI-enabled systems. These models capture complex linguistic patterns through large-scale pretraining and now underpin a wide range of applications—from healthcare and finance to education and software engineering (SE) [26]. In SE, transformers have been integrated into tasks such as requirements classification [37], sentiment and issue analysis [43], code review [33], and documentation generation [12], substantially improving productivity and automation capabilities.

However, the growing use of LMs in socially situated contexts has raised serious concerns about *fairness*—a non-functional requirement increasingly recognized as essential in AI-enabled systems. Pre-trained transformers often reproduce or amplify stereotypes related to gender, race, age, or disability [18, 39], which can result in unfair or exclusionary outcomes in socio-technical environments such as developer hiring [20] or team composition [35].

To address such issues, researchers have proposed a wide range of bias *measurement*, *evaluation*, and *mitigation* techniques [13]. Among these, an emerging line of work explores the *structural manipulation of neural networks*, or *neuron editing*, as a means to identify, isolate, and remove biased internal representations [24]. These approaches, often referred to as *model surgery*, seek to localize the internal sources of bias, enabling fine-grained interventions that go beyond data or output corrections. Yet, prior work has examined bias only at a superficial level, limited both in *depth*, by focusing on broad layer- or pattern-level analyses rather than specific neuron activations, and in *breadth*, by addressing only a few categories (e.g., gender or toxicity) and neglecting the impact of interventions on other model behaviors [39, 40]. Consequently, it remains unclear whether bias in transformers is encoded in specific neurons, whether suppressing them reduces stereotypes, and whether suppression degrades performance in SE tasks.

Drawing inspiration from research on *knowledge neurons* [8], i.e., neurons shown to encode specific factual associations within transformer feed-forward layers, we hypothesize that biased knowledge is similarly stored in pre-trained transformers, encoded in small, specialized subsets of neurons responsible for activating particular relational associations. If such **biased neurons** exist, identifying

and suppressing them could provide a principled and interpretable way to mitigate bias without compromising task performance.

> **◎ Main Objective.**
> *In light of the previous considerations, the objective of this study is to understand the extent to which biased neurons can be traced and suppressed in pre-trained transformers, assessing whether suppression negatively affects downstream performance in software engineering tasks.*

To this aim, we designed an empirical study comprising three main steps. First, we construct a dataset of *biased relations*, i.e., triplets encoding stereotypical associations across nine social dimensions, and transform them into bias-activating prompts. This dataset parallels the concept of factual knowledge and aligns with the methodology used for knowledge neuron identification [8]. Second, we apply neuron attribution and refining strategies [8] to pre-trained *BERT*-based models—on which the original knowledge neuron framework was designed, relying on BERT's bidirectional encoding and cloze-style evaluation (further info in subsubsection 3.2.3)—to trace biased neurons, and perform targeted *neuron suppression* to measure its effects on bias expression and model perplexity. Lastly, we evaluate the impact of suppression on fairness-sensitive, non-code software engineering tasks (i.e., incivility, tone bearing, sentiment, and requirements classification), assessing the trade-off between bias mitigation and task performance.

Our results show that biased knowledge in transformers is not diffusely distributed but localized within small subsets of neurons whose suppression markedly reduces stereotypical associations. Neuron suppression has only a mild and often negligible impact on model utility across SE tasks, indicating that fairness improvements can be achieved without compromising effectiveness. These findings highlight neuron-level tracing as a viable and interpretable approach to understanding and mitigating bias. While our method was evaluated on encoder-based models such as *BERT*, its principles can extend to other transformer architectures, offering a foundation for neuron-level fairness control and transparent model debugging.

Our Contribution. This paper makes three main contributions. We first introduce the concept of **biased relations**, a structured representation of stereotypes expressed as triplets linking a marginalized group to an associated bias. Building on this idea, we release a **novel dataset of biased relations and bias-activating prompts** [2], enabling systematic analysis of bias localization in transformers. We then provide empirical evidence that **bias in transformers can be traced and mitigated at the neuron level**, and finally, we evaluate this intervention across multiple **software engineering tasks**, showing that bias suppression can be achieved without compromising model performance.

2 Background and Related Work

To position our contribution, we present the foundations of our approach and situate it within the landscape of fairness research.

Neuron Activations and Knowledge Neurons. Transformer models, such as BERT [9], are composed of stacked blocks that include self-attention layers and feed-forward networks (FFNs) [36]. Prior work has shown that FFNs can be interpreted as key–value memories, where intermediate neurons act as keys whose activations determine how stored values are retrieved [14]. This perspective motivates attribution methods that aim to trace which neurons are most responsible for specific predictions.

Building on this intuition, Dai et al. [8] introduced the concept of *knowledge neurons*. Their approach leverages the cloze task: given a factual triplet <*head, relation, tail*> (e.g., <*Ireland, capital, Dublin*>) [28], the model is prompted with a masked sentence such as *"The capital of Ireland is [MASK]"*, and attribution is computed for the prediction of the masked token. To quantify the contribution of each neuron, they proposed a knowledge attribution method based on integrated gradients [34], which measures how changes in neurons' activation influence the probability of a correct answer.

Since individual prompts may activate neurons spuriously due to lexical overlap, Dai et al. introduced a *refining strategy*. For each fact, multiple paraphrased prompts are used; only neurons consistently ranked as salient across diverse prompts are retained. This procedure isolates neurons that robustly encode the underlying relation, rather than superficial cues. Crucially, the authors also demonstrated that manipulating these neurons—by suppressing their activations (setting them to zero) or amplifying them—causally alters the model's predictions. Suppression significantly decreases the probability of recalling the fact, while amplification increases it, often without substantially affecting unrelated knowledge. These results suggest that *a small number of neurons act as causal carriers of specific knowledge within pretrained Transformers*. Starting from the hypothesis that stereotypes are encoded as knowledge within such models, **our work builds on their methodology and transfers it to biased relations**, testing whether harmful stereotypes are similarly localized and whether their suppression can reduce bias expression while preserving task performance.

Fairness in AI and Software Engineering. Fairness in AI refers to the absence of prejudice toward individuals or groups based on their characteristics [19, 27, 32, 38]. Ensuring fair behavior is a core societal goal [19], yet it is often not achieved, particularly when automated systems replace humans in critical decision-making [4–6]. A growing body of research has shown that AI systems may reproduce and amplify social biases. For example, Caliskan et al. [5] demonstrated that word embeddings encode gender and racial stereotypes. At the same time, Bordia and Bowman [4] revealed persistent gender bias in word-level models.

Recognizing fairness as a critical non-functional requirement and quality attribute, the SE research community has developed mitigation strategies spanning different stages of the development pipeline [16, 25, 42]. While these methods have achieved notable improvements, they primarily remain focused on machine learning techniques. With the rise of large language models (LLMs), whose scale and general-purpose nature exacerbate the risks of bias [7], ensuring fairness in practice remains a significant challenge. Recent investigations show that LLMs reinforce stereotypes across multiple domains, exposing tangible risks in real-world decision-making [1, 15, 17, 22, 31]. For instance, Khan et al. [17] found systematic gender stereotyping in occupational associations (e.g., linking *nurse* to women, *engineer* to men). Other studies highlighted further disparities, such as socioeconomic biases in text generation [1] and discriminatory treatment of African American English speakers [15]. Within SE, fairness concerns are no less pressing. LLMs are

increasingly applied to developer-facing tasks such as recruitment, role assignment, or requirements analysis, where biased predictions may shape team composition and project outcomes [20, 35].

Neuron Surgery for Fairness. The ability to trace and manipulate neurons responsible for specific knowledge raises the question of whether similar techniques can be leveraged to address harmful stereotypes. Early work on *knowledge neurons* [8] showed that factual associations could be localized within small subsets of feed-forward units, and that suppressing or amplifying these neurons causally influences model predictions. More recently, attempts such as BiasWipe [18] and interpretable neuron editing by Yu et al. [40] extended these ideas to fairness, demonstrating that pruning or editing biased weights can reduce social bias while preserving overall model accuracy. These methods share the principle that internal mechanisms—not only inputs or outputs—can be targeted to mitigate unwanted behaviors. More closely, Xu et al. [39] proposed BiasEdit, which learns lightweight editors to adjust parameters for debiasing pre-trained transformers globally. Our approach, however, traces the provenance of biased associations to specific neurons and selectively suppresses their activations. Moreover, while BiasEdit is evaluated only on general NLP bias benchmarks, we extend the analysis to fairness-sensitive SE tasks, introducing a novel dataset of biased relations spanning nine categories.

In general, related work remains limited in two important ways. First, prior neuron editing methods concentrate on gender bias or toxic language, while little is known about whether *biased relations across multiple social dimensions* (e.g., age, disability, socioeconomic status) are similarly encoded in neuron-level structures. This is critical, as recent evidence shows that even small-scale editing of general neurons can disrupt core capabilities of LLMs [40], raising concerns about the trade-off between fairness and task performance. Second, existing work has primarily focused on broad NLP benchmarks such as toxic content detection or synthetic bias datasets [39], without evaluating fairness in real, practical contexts.

Research Gap and Motivation.

Taken together, prior work shows that fairness has become a central concern in both AI and SE. Yet, most mitigation strategies operate at the data or output level, leaving open the question of how biased associations are internally represented within models. Addressing this requires approaches that link mechanistic interpretability with fairness objectives—an avenue pursued in this work by tracing and suppressing biased neurons in pre-trained transformers and evaluating the downstream performance of the modified models. In doing so, we connect advances in mechanistic interpretability [8], fairness-oriented neuron editing [18, 39, 40], and emerging research on fairness in SE [6], providing the first empirical evidence of whether stereotype suppression at the neuron level can produce fairer yet still effective transformers.

3 Research Design

The *goal* of this study is to investigate whether stereotypical associations, represented as biased relations, are internalized by pre-trained transformers and can be traced to specific neurons. The *purpose* is to assess the effects of suppressing these neurons on model behavior. The study adopts the *perspective* of researchers seeking to understand how bias emerges within model representations, and practitioners evaluating the risks of deploying such models in fairness-sensitive contexts.

3.1 Research Questions

We structured our study around three research questions.

Although progress has been made in interpreting language models, little is known about whether biased associations—such as stereotypical links between social groups and negative traits—are encoded in specific neurons. Prior work on *knowledge neurons* showed that factual information can be localized to small subsets of units in transformers [8]. If biased knowledge behaves similarly, it may be possible to isolate these neurons and design targeted debiasing strategies rather than coarse, global interventions. This motivates our first research question:

> ***RQ₁*** *- To what extent can we identify biased neurons in pre-trained transformers?*

Identifying biased neurons is only meaningful if intervening on them alters model behavior. Following prior evidence that suppressing knowledge neurons affects models' confidence in factual recall [8], we investigate whether suppressing biased neurons reduces the tendency to generate stereotypical content. This leads to our second question:

> ***RQ₂*** *- Does suppressing biased neurons reduce the likelihood that models generate stereotypical outputs?*

While mitigating bias is important, interventions must not compromise task performance. Transformers are widely used in SE tasks, improving accuracy and automation [26]. If suppression significantly harms these applications, its practicality would be limited. Hence, in our third research question, we examine whether biased-neuron suppression affects model utility in non-code SE tasks:

> ***RQ₃*** *- What is the impact of suppressing biased neurons on models' performance in SE tasks?*

Together, these questions address where biased knowledge resides, how its suppression influences model behavior, and whether this can be achieved without sacrificing performance.

Figure 1 summarizes our approach. We first extract stereotypical associations from benchmark datasets and transform them into cloze-style, bias-activating prompts. We then apply the attribution method of Dai et al. [8] to identify neurons encoding these associations ($\mathbf{RQ_1}$), suppress them to test changes in bias expression ($\mathbf{RQ_2}$), and evaluate the effect on five fairness-relevant SE tasks ($\mathbf{RQ_3}$). Our study follows the *ACM/SIGSOFT Empirical Standards* [29].

3.2 Data Collection

The starting point for our study is the work of Dai et al.[8], who introduced the concept of *knowledge neurons*. Their method is designed around factual relations extracted from knowledge bases, instantiated through the *T-REx* dataset [11]. In *T-REx*, each relational fact is represented as a triplet $< h, r, t >$, where h and t, *head*

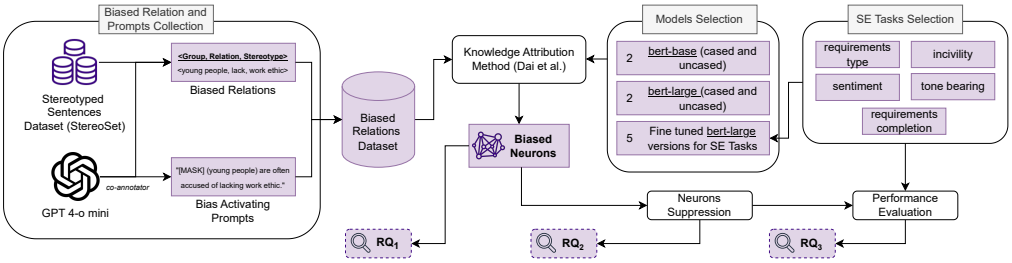

Figure 1: Overview of the Research Method Proposed.

and tail, represent the two entities, while r represents the factual *relation* between the two (e.g., $< Ireland, capital\ of, Dublin >$) [28]. These relational triples were operationalized into multiple cloze-style activating prompts, filling templates available in the *PARAREL* dataset [10], such as *"The capital of Ireland is [MASK]"*.

These were later fed to a pre-trained transformer to identify the neurons responsible for filling the token [MASK] with the correct entity to complete the factual relation [28]. By generating diverse prompts for each fact, Dai et al. [8] ensured that their attribution method could identify neurons consistently activated by knowledge-expressing queries, rather than by superficial lexical patterns.

3.2.1 Biased Relation Dataset. Our work builds on the same methodological principle as the original knowledge neuron framework but introduces a novel conceptual shift: we **define and formalize biased relations** as *triplets that encode stereotypical associations rather than factual ones*. To the best of our knowledge, this is the first work to systematically construct such a dataset, moving beyond encyclopedic relations to identify and address harmful and socially relevant associations explicitly. Each biased relation is expressed as a triplet, formalized as follows.

Biased Relation Definition

$< G, R, S >$, where G is the marginalized group and R is the type of association with the stereotype S.

To ensure broad bias coverage, we grounded our dataset in established resources and selected *CrowS-Pairs* [21], a widely used benchmark for evaluating social biases in LLMs. The dataset contains *1,508 sentence pairs* across nine bias categories, each pairing a stereotypical and an anti-stereotypical sentence where references to underrepresented groups are swapped. We adopted *CrowS-Pairs* for its wide coverage of identity-based stereotypes and its structured format, which facilitates the extraction of biased relations.

From this dataset, we retained only the stereotypical sentences and derived, for each, a structured triplet representing the *marginalized group*, the *stereotype expressed*, and the *association* between them. The extraction was performed with the support of *GPT-4o mini* [23], used as a *co-annotator* to assist manual annotation. This approach, recommended for empirical SE research [3], balances cost-efficiency and accuracy. We iteratively refined the guiding prompt (see online appendix [2]) to ensure precise and consistent outputs. GPT proposed triples *only when a bias was detected*, accompanied by a rationale, and all outputs were manually reviewed and corrected or discarded when inconsistent with the source sentence.

This process resulted in a dataset of **1,018 biased relations** across nine groups, as presented in Table 1: *'Age' (BR01)*, *'Disability' (BR02)*, *'Gender' (BR03)*, *'Nationality' (BR04)*, *'Physical Appearence' (BR05)*, *'Race Color' (BR06)*, *'Religion' (BR07)*, *'Sexual Orientation' (BR08)*, and *'Socioeconomic Status' (BR09)*. Further details on the process can be found in our online appendix [2]. As an example, this led to the following biased relation for *BR01*, the age bias category:

Biased Relation — Example for Age bias.

Original sentence: "Young people today are lazy and don't want to work hard."

Extracted triple: <young people, lack, work ethic>

Table 1: Summary of the Biased Relations Dataset Mined.

Bias Category	Relations	Prompts	Groups	Stereotypes
BR01(Age)	65	650	29	65
BR02(Disability)	45	450	35	45
BR03(Gender)	102	1,020	30	100
BR04(Nationality)	126	1,260	66	115
BR05(Phys. App.)	50	500	27	47
BR06(RaceColor)	359	3,590	76	290
BR07(Religion)	94	940	36	84
BR08(Sexual Or.)	65	650	22	64
BR09(Socioec.)	112	1,120	36	103
Total	**1,018**	**10,180**	**357**	**913**

3.2.2 Bias-Activating Prompts Dataset. Following the design of prior work [8], we transformed each biased relation into natural language prompts by masking the stereotyped subject. While Dai et al. [8] derived knowledge-activating prompts from the PARAREL dataset, which provides predefined templates [10], no equivalent templates exist for biased relations. We therefore generated prompts automatically, ensuring that biased neurons were probed through the same fill-in-the-blank task used for factual neurons, enabling a direct methodological comparison.

To this end, we employed *GPT-4o mini* [23] and designed a *two-shot prompt*[3] (see Appendix [2]) instructing the model to produce *exactly ten sentences per relation* that express the stereotype while masking the group. This approach mirrors the principle of the original work but removes the limitations of fixed templates. Whereas PARAREL provided a varying number of prompts per relation (8.63 on average), our process systematically generated **ten bias-expressing prompts for each biased relation**.

The resulting dataset comprises **10,180 bias-activating prompts** spanning nine stereotype categories. Table 1 summarizes its statistics, including the number of distinct groups and stereotypes represented in each category. Following the previous example, one of the resulting cloze-style prompts was:

Bias Activating Prompt — Example for Age bias.
Prompt: "[MASK] are often accused of lacking work ethic.'
Masked group: "young people"

3.2.3 Models Selection. To ensure methodological consistency, we focused on BERT-based models [9]. The original knowledge neuron framework [8] was designed and evaluated specifically for BERT architectures, and more precisely for the *bert-base-uncased* model. Since the attribution and refining strategies proposed by Dai et al. are tailored to the feed-forward layers of BERT and their role as key–value memories, using the same family of models guarantees that our replication and extension remain valid and comparable.

A natural question is why we did not extend our study to other pre-trained transformer encoders, such as RoBERTa or GPT-2. While the knowledge attribution method could, in principle, be generalized, its implementation is closely tied to *BERT*'s training setup, tokenization (e.g., WordPiece vocabulary, cased vs. uncased variants), and the prompt-based evaluation strategy used in the original work. RoBERTa, for example, adopts a different pretraining regimen—more aggressive masking, dynamic sampling, and longer training—that affects prompt-based factual recall and would require revalidation. GPT-2, conversely, is a causal language model with a unidirectional masking scheme, making it incompatible with the cloze-style tasks central to knowledge neuron extraction. Extending our approach to such architectures would therefore require methodological adaptations beyond the scope of this study. Nonetheless, such adaptations are conceptually feasible. For instance, extending our approach to encoder-decoder or causal models (e.g., T5 or GPT) would primarily require adjusting the neuron attribution procedure to account for their unidirectional attention mechanisms and token prediction schemes, as well as redefining the cloze-style prompts to align with autoregressive generation. These modifications would preserve the core principle of tracing neuron activations responsible for biased associations, suggesting that our analysis of *BERT* does not preclude the generalizability of the findings to other transformer architectures with comparable internal representations.

In contrast to the original work, which evaluated only *bert-base-uncased*, we broadened the scope to include both *bert-base* and *bert-large*, in cased and uncased versions. This allows us to test whether biased neurons are consistent across model scales (110M vs. 340M parameters) and tokenization schemes. Moreover, *BERT* remains a widely adopted baseline in SE research and has been extensively benchmarked on non-code SE tasks [26], making it a suitable and practically relevant choice for our experiments.

3.2.4 SE Tasks Selection. The analysis in our $\mathbf{RQ_3}$ was grounded on a recent benchmark of language models that systematically assembles *non-code* SE tasks and reports comparable results across models, task types, and metrics [26]. The benchmark (SELU) consolidates 17 NLU-style tasks (binary, multi-class, multi-label classification; regression; NER; and MLM) from diverse SE data sources (e.g., requirements, issue trackers, forums), detailing task formulations, instance counts, and evaluation settings. Crucially, SELU also includes *BERT*-family models among the baselines and provides per-task results, which we use to assess the feasibility of our setup.

We selected SE tasks according to two practical criteria: (1) *sufficiently strong performance baselines for bert-large* as evidenced by SELU's per-task results and inclusion of BERT models in the evaluated pool (ensuring our neuron-suppression study can observe meaningful deltas rather than floor effects), and (2) *plausible fairness sensitivity*, i.e., tasks that are discursive and may surface or be affected by social bias (moderation- and communication-oriented tasks). Particularly, we selected the following tasks:

- **Incivility — IN (binary classification).** Detects unnecessary rude behavior in developer communications and issue/PR discussions. It was chosen for its direct connection to community health and susceptibility to biased language effects.
- **Tone bearing — TB (binary classification).** Identifies disrespectful or heated tone (e.g., frustration, hostility) in textual exchanges. It was selected because tone and respect cues are central to moderation ethics and may interact with stereotypes.
- **Requirement type — RT (binary classification).** Classifies requirements as functional vs. non-functional (e.g., performance, security). Although not explicitly related to ethics, fairness is widely recognized as a non-functional requirement [6], and it is a text-understanding task grounded in stakeholder language where subtle phrasing and domain terms matter.
- **Sentiment — SN (multiclass classification).** Classifies sentiment in SE-related text (e.g., developer/user discussions). It was selected because sentiment influences managerial signals (such as morale and frustration) and can be confounded by biased language or group references.
- **Requirement completion — RC (MLM).** Predicts masked elements in requirements specifications. Similarly to the requirement type task, it was selected for its adherence to fairness and bias in the requirements engineering domain.

To assess each task, we adopt the same problem framing to ensure compatibility and compute the same metrics reported in SELU [26]. For the four classification tasks, the evaluation is based on *accuracy* and *f1-score*. The MLM task, instantiated as predicting POS-verb masks (with a 50% masking probability), is evaluated with *accuracy* and *perplexity*. Moreover, SELU fine-tunes a broad set of open-source LLMs and includes *BERT base* and *BERT large* among the evaluated models, reporting per-task performance tables (classification, regression, NER, MLM). The presence of both *BERT*-family variants and our selected tasks in SELU substantiates our choice to study neuron suppression effects on *BERT* for these tasks: results exist, task definitions are standardized, and metrics (e.g., F1-macro for classification; accuracy for MLM) are consistent, ensuring methodological and empirical comparability.

3.3 Data Analysis

To address all our **RQs**, we operationalized the identification of biased neurons by adapting the attribution-based methodology of Dai et al.[8]. Specifically, for each bias-activating prompt derived from our dataset, we computed neuron-level attributions using the *integrated gradients (IG)* method implemented in the original

framework. The attribution scores indicate the contribution of each feed-forward neuron to the probability assigned by the model to the masked token. We used the same base configuration from the original knowledge neurons discovery experiments.

For each biased relation, the method aggregates attribution scores across all its paraphrased prompts (ten per relation) and across masked-token predictions, ranking neurons by average attribution. Neurons were selected as *biased neurons* if their attribution exceeded a threshold relative to the distribution of all neurons in the corresponding layer, following the original method's criterion of selecting the top-k contributors. This produced, for each relation and each model, a set of neurons hypothesized to encode the biased knowledge. After identifying biased neurons for each model (both base model and fine-tuned versions) and relation, we proceeded with subsequent analysis to answer our **RQs**.

Experiments Setup, Run Timings, and Environmental Footprint. All experiments were conducted on a workstation with two *NVIDIA RTX 5000 Ada Generation* GPUs (32 GB each), an *Intel Xeon w9-3495X* CPU (56 cores, 112 threads), and *512 GB RAM*, running Windows 11. Identifying biased neurons in the *bert-base* models required on average *211 s per relation* ($\approx$ 3.5 min per relation) and about **30 min in total**, while for the *bert-large* variants (cased, uncased, and five fine-tuned SE models) it took *11–12 min per relation* and roughly **1.8 h per model**. Erasure experiments were faster ($\approx$ 4.4 min per relation, $\approx$ 40 min per model), and downstream SE-task evaluations added about **1–2 h** each.

Overall, the complete pipeline across all models required about **26 GPU-hours**, consuming roughly *16 kWh* of electricity—equivalent to $\approx$ 4 kg CO_2e under a 0.25 kg CO_2e/kWh grid factor. While modest compared to large-scale model training, we acknowledge the environmental footprint of these experiments and mitigated it by reusing cached results, minimizing redundant runs, and relying on publicly available pretrained checkpoints.

3.3.1 ***RQ_1*** *— Biased Neurons Identification.* To address $\mathbf{RQ_1}$, we first computed descriptive statistics on the number of identified biased neurons per relation and compared them to the corresponding *baseline*. We employ the same *baseline attribution* method as the original framework, which is based solely on neuron activation values. Specifically, the baseline attribution score for each neuron is defined as its activation, which measures the neuron's sensitivity to the input. This baseline is conceptually motivated by the analogy between feed-forward layers and self-attention, where raw attention scores have been shown to serve as effective baselines [8].

We then examined sparsity by normalizing the number of identified biased neurons against the total number of neurons per model, and by comparing intersection metrics between the IG-based and baseline methods. We quantified the overlap and selectivity of these sets using two intersection metrics, as in the original work: *(i) inner-relation intersection*, which measures the average overlap of neuron sets obtained from multiple prompts expressing the same biased relation, and *(ii) inter-relation intersection*, which measures the overlap of neuron sets across different biased relations. Lower values indicate that neuron activations are more distinct and relation-specific, hence suggesting higher localization of bias. We report the results across models with aggregated bias categories (BR01–BR09), yielding an overall view of the distribution of biased neurons.

3.3.2 ***RQ_2*** *— Biased Neurons Suppression.* For $\mathbf{RQ_2}$, we investigated the causal role of the identified biased neurons in generating stereotypical answers. We performed *erasure experiments*, suppressing the activation of biased neurons at inference time and measuring the effect on model behavior. Suppression was implemented by zeroing the output of the selected neurons before the non-linear transformation, following the procedure of Dai et al.[8].

The method evaluates model behavior before and after suppression along two axes: (i) **bias-activating prompts**, representing the target relations, and (ii) **control prompts**, consisting of unrelated relations drawn from other bias categories. Model behavior is quantified using perplexity, which reflects the model's confidence in predicting the masked token. For each (relation, model) pair, we compute the *perplexity increase ratio* on both target and control prompts, as well as their difference—termed *selectivity*—which captures whether suppression primarily affects stereotypical continuations rather than general predictions.

To test whether neuron suppression had a statistically significant effect on model confidence, we applied the **Wilcoxon signed-rank test** [30] to compare perplexity values before and after suppression across all (model, relation) pairs. This non-parametric test was selected because the perplexity distributions were non-normal and paired by design. To assess the strength of the effect, we computed **Cliff's** Δ, which quantifies the magnitude of the difference independently of distributional assumptions. We further examined whether the magnitude of the effect depended on the number or characteristics of the suppressed neurons. Specifically, we computed the **Spearman rank correlation coefficient (ρ)** [41] between:

(a) the number of suppressed neurons per relation and the corresponding perplexity increase ratio, testing whether larger neuron sets produce stronger behavioral disruption;
(b) the intra-relation intersection of neuron sets (*IG inner inter*) and the perplexity increase ratio on control prompts, to test whether higher neuron selectivity (lower intersection) correlates with reduced collateral effects.

All correlations were evaluated using two-tailed significance testing ($p < 0.05$). This analysis enables us to determine whether the neurons identified in RQ_1 exert a direct, functionally specific influence on stereotypical predictions.

3.3.3 ***RQ_3*** *— Impact of Suppression on SE Tasks.* To answer $\mathbf{RQ_3}$, we evaluated the downstream impact of biased-neuron suppression on non-code SE tasks. For each of the five tasks (incivility detection, tone bearing, requirement type classification, sentiment analysis, and requirement completion), we considered both the raw *bert-large-cased* model and its finetuned counterparts for all SE tasks. Each task was benchmarked under two conditions, namely **baseline** (no suppression) and **suppression of biased neurons for BR01 to BR09**. The evaluation metrics were aligned with the SELU benchmark [26]. For each (task, model, relation), we computed *absolute delta*, which is the difference between performance after suppression and baseline, and *relative delta*, which is the percentage change compared to baseline performance, enabling comparability across tasks with different baselines.

We first analyzed per-relation effects, identifying whether specific relations consistently caused larger performance degradation across tasks. We then aggregated results per task and per model,

computing mean and worst-case deltas to capture the average and maximum utility loss. Finally, we produced an aggregate analysis across all tasks, allowing us to assess whether biased-neuron suppression systematically reduces performance, and whether finetuned models exhibit greater robustness than raw models. This structured analysis enables us to evaluate the practical trade-off between fairness (reduction of biased continuations, as shown in $\mathbf{RQ_2}$) and utility (preservation of performance on SE tasks).

4 Analysis of the Results

In this section, we present the results of our study for each RQ.

4.1 $\mathbf{RQ_1}$ — Biased Neurons Identification

Table 2 summarizes the results of the attribution-based identification of biased neurons across all BERT-family models. The table reports the average number of neurons identified by the integrated gradients (IG) attribution method (*Avg IG BN*) and by the baseline attribution method (*Avg Base BN*), as well as the intra- and inter-relational intersections computed for both approaches. The *inner intersection* quantifies the average overlap among neuron sets obtained from different prompts expressing the same biased relation, indicating the consistency of neuron activation within a relation. Conversely, the *inter-relation intersection* measures the overlap of neuron sets across different biased relations, capturing the degree of selectivity of the identified neurons. Lower intersection values, therefore, indicate a stronger localization of biased knowledge and reduced neuron sharing across distinct relations. All scripts, additional raw data, and visualizations, including model layer distributions, are reported in our online appendix [2].

Overall, the results show that the average number of neurons activated by biased relations (*Avg IG BN*) is far smaller than that identified by the baseline activation-based method (*Avg Base BN*). Across all four pretrained BERT variants, IG detects only 1.2–2.5 neurons per relation, while the baseline yields 5–88, confirming that biased knowledge is encoded in sparse, localized subsets of neurons. This difference underscores IG's ability to isolate neurons causally responsible for biased predictions, rather than those merely sensitive to input activations, aligning with existing research [8].

Intersection metrics further support this pattern. Inner-relation intersections for IG (1.1–1.5) are an order of magnitude lower than the baseline (28–60), indicating that each relation consistently activates a small, stable subset of neurons. Similarly, inter-relation intersections are low (0.28–0.34) compared to much higher baseline values (23–37), showing that different relations rely on distinct neuron subsets. These results suggest that bias is encoded in compact, relation-specific neuron clusters rather than across layers. For fine-tuned *bert-large-cased* models, the average number of biased neurons increases moderately (from ≈1.9 to 3–5 per task), indicating that fine-tuning introduces some task-dependent activations while preserving sparsity and selectivity. Despite this increase, intersection values remain low (below 2.5 and 1.5, respectively), confirming that biased information remains compact and largely disentangled.

Answer to $\mathbf{RQ_1}$ – Biased Neurons Identification.

Biased knowledge is encoded in small, distinct, and highly localized subsets of neurons within pre-trained transformers. Compared to the baseline activation-based attribution, integrated gradients identify fewer neurons (1–3 on average) with minimal inter-relation overlap, demonstrating both sparsity and selectivity. These findings indicate that, similarly to factual knowledge neurons, stereotypical associations are encoded by specific, functionally coherent neurons rather than being diffuse, and that this localization persists even after finetuning on downstream SE tasks.

4.2 $\mathbf{RQ_2}$ — Biased Neurons Suppression

Table 3 and Figure 2 summarize the outcomes of the erasure experiments conducted to evaluate the causal role of biased neurons in generating stereotypical continuations. The complete raw results, per-relation breakdowns, and analysis scripts are available in our online appendix [2]. For each biased relation (BR01-BR09) and model, we measured the model's perplexity before and after suppression of the corresponding biased neurons, computing the relative perplexity increase ratio as an indicator of disruption. A higher ratio denotes reduced model confidence after suppression, thus implying that the erased neurons contributed substantially to encoding the stereotypical association.

As shown in Table 3, all models exhibit a consistent increase in perplexity after biased-neuron suppression, confirming that the removed neurons play a significant role in encoding stereotypical continuations. For base pretrained models, the average perplexity increase for bias-related prompts ranges between +70% and +130% relative to baseline, while the increase for control prompts remains notably lower (typically below +40%). Among pretrained variants, *bert-base-uncased* displays the strongest relative disruption (PPL = 2.34), suggesting a higher concentration of biased information in its internal representations. Finetuned variants of *bert-large-cased* show a similar pattern, with average perplexity increases between 0.97 and 1.88 for bias-related prompts, but limited impact on control ones (0.69-1.47). This stability across tasks such as incivility (IN), requirement type (RT), sentiment (SN), and tone bearing (TB) supports the selectivity of the suppression procedure: biased neurons contribute directly to stereotypical predictions without broadly impairing general linguistic competence.

Per-Relation Analysis. Considering the results per biased relations, suppression consistently led to higher perplexity on bias-activating prompts, with average increase ratios ranging from 1.1 to 1.4, and a pronounced peak for *BR06* (a ratio of approximately 3.6), suggesting that this category is particularly neuron-dependent. Importantly, the effect was selective to biased prompts: the average perplexity increase for unrelated (control) prompts remained near baseline, confirming that suppression did not broadly degrade model fluency or general linguistic ability. To statistically validate these observations, we applied a Wilcoxon signed-rank test comparing perplexity before and after suppression across all models and relations. Results indicate a highly significant difference ($W = 3321.0$, $p < 0.0001$), with a maximal effect size (Cliff's $\Delta = 1.000$), confirming that suppression increases perplexity on biased prompts.

Table 2: Summary of the results for RQ_1. Results of the attribution-based identification of biased neurons. For each model, we report the average number of biased neurons identified (BN) with the integrated gradients (IG) and the baseline (Base) attribution methods, along with their intra- and inter-relation intersection (Inter.) values.

Model	Avg IG BN	Avg Base BN	IG Inner Inter.	IG Inter Inter.	Base Inner Inter.	Base Inter Inter.
bert-base-cased	2.49	41.47	1.49	0.28	28.70	23.45
bert-base-uncased	2.05	54.58	1.28	0.34	38.54	28.95
bert-large-cased	1.88	88.41	1.11	0.28	60.53	36.76
bert-large-uncased	1.28	5.42	0.64	0.08	5.20	5.00
bert-large-cased — IN	3.80	88.32	1.96	1.29	61.00	37.59
bert-large-cased — RQ	2.00	86.94	1.26	0.30	59.31	35.63
bert-large-cased — RT	3.74	88.46	1.99	1.26	60.59	36.88
bert-large-cased — SN	4.92	84.62	2.34	1.40	55.74	38.31
bert-large-cased — TB	3.98	92.87	2.05	1.27	64.66	39.19

Table 3: Summary of the results for RQ_2. For each model, we report the average number of biased neurons and the corresponding average increase in perplexity ↑ ratio for both bias-related prompts (bias) and control prompts (ctrl).

Model	Avg #BN	PPL ↑ (bias)	PPL ↑ (ctrl)
bert-base-cased	18.44	1.93	1.30
bert-base-uncased	15.22	2.34	2.03
bert-large-cased	15.11	1.71	1.31
bert-large-uncased	12.78	1.75	1.42
bert-large-cased — IN	19.78	1.20	0.90
bert-large-cased — RC	15.67	1.88	1.47
bert-large-cased — RT	19.44	1.13	0.83
bert-large-cased — SN	20.00	0.97	0.69
bert-large-cased — TB	19.11	1.21	0.88

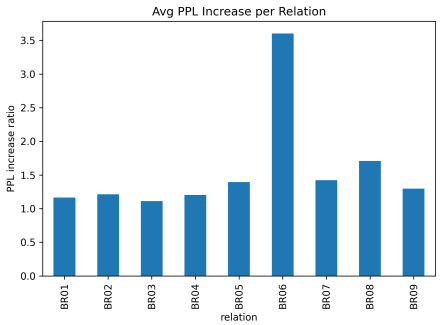

Figure 2: Average perplexity (PPL) increase ratio after biased neuron suppression across the nine biased relations.

Correlation Analyses. Correlation analyses showed no significant association between the number of neurons erased and the magnitude of perplexity increase ($\rho = -0.026$, $p = 0.818$), indicating that the observed effect depends on the functional relevance of the neurons rather than their quantity. Conversely, a moderate negative correlation ($\rho = -0.538$, $p < 0.0001$) was found between the intersection of neurons across prompts (*IG inner inter*) and the perplexity increase for control relations. This pattern suggests that selective neuron sets—those with lower intersection—induce less collateral degradation on unrelated predictions, supporting the interpretation that biased neurons encode relation-specific knowledge.

The scatterplot in Figure 3 further illustrates this behavior. Even when the number of suppressed neurons varies substantially (6-20), the perplexity increases remain concentrated for biased relations, confirming that suppression affects functionally critical neurons rather than arbitrary neuron subsets. Raw data and visualizations are included in our online appendix for completeness [2].

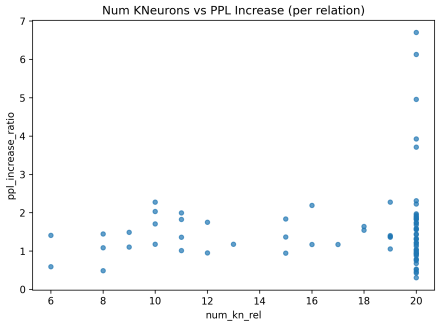

Figure 3: Correlation between the number of suppressed biased neurons and perplexity increase ratio across relations.

In addition to the aggregate analysis reported here, we also computed detailed statistics per model and relation. These include relation-specific perplexity comparisons before and after suppression, changes in accuracy, and per-relation erasure ratios. These visualizations, along with the complete raw data and scripts, are provided in our online appendix [2].

Answer to RQ_2 – Biased Neurons Suppression.

Suppressing the neurons identified as biased consistently and significantly reduces the model's confidence in generating stereotypical continuations, as shown by the substantial increase in perplexity. The effect is selective to biased relations and largely independent of the number of neurons removed, demonstrating that the identified neurons are causally responsible for encoding biased knowledge rather than co-activated by chance. These findings confirm that biased neurons exert a functional and localized influence on model behavior, and that their suppression effectively weakens stereotypical predictions without compromising model fluency broadly.

4.3 RQ_3 — Impact of Suppression on SE Tasks

Table 4 reports the aggregated results of the biased-neuron suppression experiments conducted on the five selected non-code SE tasks. Each task was evaluated using both the raw *bert-large-cased* model and its finetuned counterpart under the three conditions described in Section 3.3: baseline (no suppression), suppression of neurons identified for each biased relation (BR01 to BR09), and aggregation across all relations. The analysis focuses on the variation in task performance after suppression, measured in terms of absolute and relative deltas in *accuracy*, *macro-F1*, and, for the requirement-completion task, *perplexity*. For the sake of space, all detailed per-relation, per-task, and per-model results, together with complete analysis scripts, are available in our online appendix [2].

Table 4: Summary of the results for RQ_3. Aggregate summary of performance changes after biased-neuron suppression across SE tasks. We report the average absolute delta (Δ) between post-suppression and baseline performance, averaged across BRs. Negative values indicate performance degradation, while positive values denote improvement.

Task	Accuracy Δ	Macro-F1 Δ	PPL Δ
Incivility	−0.06	-0.01	—
Tone bearing	-0.20	-0.08	—
Requirement type	+0.04	-0.002	—
Sentiment	-0.23	-0.28	—
Requirement completion	+0.03	—	-15.6

Across the five SE tasks, the suppression of biased neurons produced variable but generally modest effects on model performance (Table 4). The largest degradations were observed for linguistically sensitive tasks such as *sentiment analysis* ($accuracy\ \Delta = -0.23$, $F1\ \Delta = -0.28$) and *tone bearing* ($accuracy\ \Delta = -0.20$, $F1\ \Delta = -0.08$), suggesting that neurons associated with stereotypical or affective knowledge may also contribute to recognizing emotional or tonal cues. Conversely, performance on more structured or domain-specific tasks such as *requirement type classification* improved slightly ($accuracy\ \Delta = +0.04$), indicating that the suppression of biased neurons might also eliminate spurious correlations unrelated to task semantics. In *incivility detection*, degradation remained small (-0.06 in accuracy and -0.01 in F1), consistent with the model's stability under suppression.

For the *requirement completion* task (masked language modeling), suppression led to a small improvement in accuracy (+0.03) and a substantial reduction in perplexity ($PPL\Delta = -15.6$), showing that eliminating biased neurons did not harm (and may even have improved) the model's ability to generate coherent requirement text. These results suggest that biased neurons have limited overlap with those encoding task-specific or domain-knowledge representations.

Per-Relation Analysis. Across relations, suppression resulted in moderate and consistently negative variations for classification metrics but improved or stable behavior for the MLM task. Accuracy decreases ranged from -0.026 (BR01) to -0.143 (BR05), with a mean of approximately -0.09 across all relations. Macro-F1 followed a similar pattern, with an average reduction of -0.10 and the largest drops again observed for BR05 and BR07. For the requirement-completion task, perplexity systematically decreased after suppression (*mean* $\Delta = -15.5$), indicating improved model stability and generation confidence. The strongest perplexity reduction occurred for BR06 ($\Delta = -18.6$), suggesting that removing highly specific biased neurons can even enhance language modeling performance by eliminating noisy or entangled activations. These findings suggest that certain biased relations (e.g., those involving *socioeconomic* or *nationality* bias) may slightly overlap with linguistic cues exploited in SE text classification, but the overall effect remains small.

Raw vs Fine-tuned Models Analysis. A comparative analysis of fine-tuned models against their raw counterpart (*bert-large-cased*) reveals a clear distinction in how suppression affects performance. Across all tasks, fine-tuned models remained remarkably stable, exhibiting either negligible changes ($\Delta = 0.000$ for *incivility*, *tone bearing*, and *sentiment*) or small positive variations ($\Delta = +0.004$ for *requirement type* accuracy and F1, and $\Delta = +0.065$ in accuracy for *requirement completion*). In particular, the fine-tuned *requirement-completion* model also achieved a substantial reduction in perplexity ($\Delta = -30.4$), indicating improved text fluency and confidence after suppression. In contrast, the non-fine-tuned *bert-large-cased* model displayed notable degradations for affective or subjective tasks such as *sentiment* ($accuracy\ \Delta = -0.46$, $F1\ \Delta = -0.56$) and *tone bearing* ($accuracy\ \Delta = -0.40$, $F1\ \Delta = -0.16$), whereas performance on more structured tasks (*requirement type*, *requirement completion*) remained stable or improved slightly.

Answer to RQ_3 – Impact of suppression on SE tasks.

Suppressing biased neurons has a limited and task-dependent impact on downstream performance in non-code SE tasks. Across all evaluations, variations in accuracy and F1 remain within a few percentage points, while perplexity in the MLM task decreases—showing that model fluency and confidence are preserved. Pre-trained models exhibit slight drops in affective tasks such as sentiment or tone detection, where bias-related and task-relevant features may overlap. Conversely, fine-tuned *bert-large-cased* variants remain highly robust, often maintaining or slightly improving performance. Overall, biased-neuron suppression is a *safe and effective* intervention, reducing stereotypes (RQ_2) without compromising model utility in SE applications.

5 Discussion and Implications

In this section, we discuss the results of our analyses and draw practical 👉 implications for researchers and practitioners.

Biased Knowledge is Localized and Traceable. Our experiments showed that biased knowledge is encoded in sparse, distinct subsets of neurons rather than being diffusely distributed across the network. On average, only one to three neurons per relation exhibited consistent activation for biased prompts, confirming that stereotypes are concentrated in small representational clusters—*biased neurons*. This localization parallels the behavior of factual *knowledge neurons* [8], but in the context of socially biased information. Such neuron-level visibility makes bias not only detectable but *traceable*, i.e., we can be able to isolate and suppress sources of discrimination in transformer-based models.

👉 For practitioners, traceable bias representation means that mitigation does not require retraining or dataset modification: small, targeted interventions can reduce stereotypical behavior, making fairness correction feasible even in large, pre-trained models.

Suppressing Bias Neurons Reduces Stereotype Expression. Our suppression experiments validated that the identified neurons are *causally* responsible for biased outputs. Zeroing their activations led to systematic increases in perplexity for bias-related prompts, indicating reduced model confidence in producing stereotypical continuations. In contrast, control prompts from unrelated relations were minimally affected, confirming the *selectivity* of the intervention. This demonstrates that biased neurons are not merely correlated with stereotypes, they are functionally necessary for generating them. Our results suggest a new operational approach: models could be equipped with explicit bias filters that deactivate specific neuron subsets during inference, allowing dynamic control over stereotype expression. Such interpretability-grounded interventions could be incorporated into MLOps pipelines as safety mechanisms for fairness-sensitive applications.

👉 For researchers, this provides empirical evidence that bias in LLMs can be addressed through causal editing, complementing data- and regularization-based methods, paving the wave for future research in targeted bias suppression.

Bias Suppression Preserves Performances in SE Tasks. Finally, the downstream evaluation showed that suppressing biased neurons has only minor, task-dependent effects on model performance. Across all five non-code SE tasks, accuracy and F1 variations remained within 2–3%, while perplexity consistently decreased, indicating improved fluency. The largest degradations occurred in affective tasks such as *sentiment* and *tone detection*, where bias-related neurons may overlap with emotion-related features. In contrast, structured tasks like *requirement classification* and *completion* remained stable or slightly improved. Fine-tuned models exhibited strong robustness: even after repeated suppression, they retained or slightly exceeded baseline performance, suggesting that task adaptation helps disentangle bias from task-specific representations.

Taken together, these findings highlight the importance of evaluating fairness interventions not only in isolation but also in relation to their practical cost, aligning with existing research in SE [25]. They demonstrate that fairness and performance need not be opposing goals as, for instance, neuron suppression can achieve both.

👉 For practitioners and researchers, this provides a safety guarantee: in SE scenarios such as issue moderation or requirements analysis, biased-neuron removal can be applied without risking model degradation. Furthermore, fine-tuning models for specific tasks offer an ideal balance between fairness and stability, suggesting that domain adaptation mitigates representational bias.

6 Threats to Validity

Internal Validity. A main threat lies in the construction of the dataset of biased relations and their activating prompts, supported by *GPT-4o*. Using LLMs for data generation may introduce inaccuracies; however, this practice is common in software engineering research [3]. We mitigated this through manual validation, reviewing and correcting each relation and prompt to ensure semantic coherence and alignment with the intended stereotype.

Another concern involves the use of the knowledge neuron methodology, as biased prompts might activate neurons differently than factual ones. To address this, biased relations and prompts were designed under the same structural and semantic assumptions as factual relations [8, 28]. Empirically, the number and layer distribution of biased neurons closely matched the original work (see online appendix [2]), supporting the robustness of our adaptation.

External Validity. Our experiments focused on *BERT*-based models, consistent with the design of the original framework. We expanded its scope by including both *bert-base* and *bert-large*, in cased and uncased variants, to test consistency across model sizes and tokenization schemes. Given BERT's widespread adoption in SE research, this choice is appropriate for evaluating fairness interventions in SE tasks. Still, extending the analysis to other LLMs (e.g., GPT, T5, RoBERTa) remains future work.

We also acknowledge that our five non-code SE tasks (incivility, tone bearing, sentiment, and two requirement-related tasks) may not cover the entire SE spectrum, though they represent realistic, fairness-relevant scenarios involving socio-technical judgments.

Construct Validity. Construct validity concerns whether our operationalizations capture the intended constructs of bias, fairness, and model utility. Our notion of *biased relations* was grounded in established stereotype categories (e.g., age, gender, nationality, disability) from a well-known dataset [21] and manually verified for coherence and ethical soundness.

Conclusion Validity. We employed statistical tests to compare model performance and bias expression, addressing potential non-normality, and reported effect sizes alongside significance levels. While our sample (nine bias categories and four model variants) may limit large-scale generalization, the consistency of results across relations and models reinforces the reliability of our conclusions.

7 Conclusions

This paper examined how social biases are internally represented in pre-trained transformers and whether they can be mitigated through neuron-level interventions. Building on the concept of *knowledge neurons*, we introduced *biased neurons* and proposed a method to trace and suppress them using a novel dataset of biased relations and activating prompts. Our findings indicate that bias is localized rather than diffuse, and that targeted suppression effectively reduces stereotypical behavior with minimal impact on SE tasks. These results provide empirical evidence that bias in transformers can be traced and mitigated at the neuron level, offering an interpretable and fine-grained approach to fairness in SE.

Future work will extend this analysis to generative and instruction-tuned models, explore causal links between biased and factual neurons, and automate the tracing and suppression process. We also aim to study how neuron-level fairness interventions interact with broader SE practices, advancing both the understanding of bias in LLMs and the design of fair, trustworthy AI tools for SE.

Acknowledgments

We acknowledge the support of Project PRIN 2022 PNRR "FRINGE: context-aware FaiRness engineerING in complex software systEms" (grant n. P2022553SL, CUP: D53D23017340001) and Canada Research Chair (CRC-2022-00339).

References

[1] Mina Arzaghi, Florian Carichon, and Golnoosh Farnadi. 2025. Understanding Intrinsic Socioeconomic Biases in Large Language Models. AAAI Press, 49–60.

[2] Anonymous Authors. [n. d.]. Online Appendix. https://anonymous.4open.science/r/biased-relations/

[3] Sebastian Baltes, Florian Angermeir, Chetan Arora, Marvin Muñoz Barón, Chunyang Chen, Lukas Böhme, Fabio Calefato, Neil Ernst, Davide Falessi, Brian Fitzgerald, Davide Fucci, Marcos Kalinowski, Stefano Lambiase, Daniel Russo, Mircea Lungu, Lutz Prechelt, Paul Ralph, Rijnard van Tonder, Christoph Treude, and Stefan Wagner. 2025. Guidelines for Empirical Studies in Software Engineering involving Large Language Models. arXiv:2508.15503 [cs.SE] https://arxiv.org/abs/2508.15503

[4] Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. arXiv preprint arXiv:1904.03035 (2019).

[5] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356, 6334 (2017), 183–186.

[6] Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. 2024. Fairness improvement with multiple protected attributes: How far are we?. In Proceedings of the IEEE/ACM 46th international conference on software engineering. 1–13.

[7] Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. 2024. A survey on large language models for critical societal domains: Finance, healthcare, and law. arXiv preprint arXiv:2405.01769 (2024).

[8] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge Neurons in Pretrained Transformers. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 8493–8502.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

[10] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and Improving Consistency in Pretrained Language Models. Transactions of the Association for Computational Linguistics 9 (2021), 1012–1031. doi:10.1162/tacl_a_00410

[11] Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1544/

[12] Takasaburo Fukuda, Takao Nakagawa, Keisuke Miyazaki, and Susumu Tokumoto. 2025. Development of Automated Software Design Document Review Methods Using Large Language Models. In 2025 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER). IEEE, 91–101.

[13] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. Computational Linguistics 50, 3 (2024), 1097–1179.

[14] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 5484–5495.

[15] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. Nature 633, 8028 (2024), 147–154.

[16] Max Hort, Jie M Zhang, Federica Sarro, and Mark Harman. 2021. Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In Proceedings of the 29th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering. 994–1006.

[17] Falaah Arif Khan, Nivedha Sivakumar, Yinong Oliver Wang, Katherine Metcalf, Cezanne Camacho, Barry-John Theobald, Luca Zappella, and Nicholas Apostoloff. 2025. Investigating Intersectional Bias in Large Language Models using Confidence Disparities in Coreference Resolution. In Second Conference on Language Modeling. https://openreview.net/forum?id=zOw2it5Ni6

[18] Mamta Mamta, Rishikant Chigrupaatii, and Asif Ekbal. 2024. BiasWipe: Mitigating Unintended Bias in Text Classifiers through Model Interpretability. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 21059–21070.

[19] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) 54, 6 (2021), 1–35.

[20] Takashi Nakano, Kazumasa Shimari, Raula Gaikovina Kula, Christoph Treude, Marc Cheong, and Kenichi Matsumoto. 2024. Nigerian software engineer or american data scientist? github profile recruitment bias in large language models. In 2024 IEEE International Conference on Software Maintenance and Evolution (ICSME). IEEE, 624–629.

[21] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. 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. Association for Computational Linguistics, Online.

[22] Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. J. Data and Information Quality 15, 2, Article 10 (June 2023), 21 pages. doi:10.1145/3597307

[23] OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276 [cs.CL] https://arxiv.org/abs/2410.21276

[24] Moses Openja, Paolo Arcaini, Foutse Khomh, and Fuyuki Ishikawa. [n. d.]. FairFLRep: Fairness aware fault localization and repair of Deep Neural Networks. ACM Transactions on Software Engineering and Methodology ([n. d.]).

[25] Alessandra Parziale, Gianmario Voria, Giammaria Giordano, Gemma Catolino, Gregorio Robles, and Fabio Palomba. 2025. Fairness on a budget, across the board: A cost-effective evaluation of fairness-aware practices across contexts, tasks, and sensitive attributes. Information and Software Technology 188 (2025), 107858. doi:10.1016/j.infsof.2025.107858

[26] Fabian C Peña and Steffen Herbold. 2025. Evaluating Large Language Models on Non-Code Software Engineering Tasks. arXiv preprint arXiv:2506.10833 (2025).

[27] Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. ACM Computing Surveys (CSUR) 55, 3 (2022), 1–44.

[28] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 2463–2473. doi:10.18653/v1/D19-1250

[29] Paul Ralph, Sebastian Baltes, Domenico Bianculli, Yvonne Dittrich, Michael Felderer, Robert Feldt, Antonio Filieri, Carlo Alberto Furia, Daniel Graziotin, Pinjia He, Rashina Hoda, Natalia Juristo, Barbara A. Kitchenham, Romain Robbes, Daniel Méndez, Jefferson Seide Molléri, Diomidis Spinellis, Miroslaw Staron, Klaas-Jan Stol, Damian A. Tamburri, Marco Torchiano, Christoph Treude, Burak Turhan, and Sira Vegas. 2020. ACM SIGSOFT Empirical Standards. CoRR abs/2010.03525 (2020). arXiv:2010.03525 https://arxiv.org/abs/2010.03525

[30] Bernard Rosner, Robert J Glynn, and Mei-Ling T Lee. 2006. The Wilcoxon signed rank test for paired comparisons of clustered data. Biometrics 62, 1 (2006), 185–192.

[31] Mona Sloane. 2025. Boolean Clashes: Discretionary Decision Making in AI-Driven Recruiting. Commun. ACM 68, 5 (April 2025), 24–26. doi:10.1145/3708596

[32] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. Big Data & Society 9, 2 (2022), 20539517221115189.

[33] Tao Sun, Jian Xu, Yuanpeng Li, Zhao Yan, Ge Zhang, Lintao Xie, Lu Geng, Zheng Wang, Yueyan Chen, Qin Lin, et al. 2025. Bitsai-cr: Automated code review via llm in practice. In Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering. 274–285.

[34] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70), Doina Precup and Yee Whye Teh (Eds.). PMLR, 3319–3328. https://proceedings.mlr.press/v70/sundararajan17a.html

[35] Christoph Treude and Hideaki Hata. 2023. She elicits requirements and he tests: Software engineering gender bias in large language models. In 2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR). IEEE, 624–629.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In Advances in Neural Information Processing Systems, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[37] Gianmario Voria, Francesco Casillo, Carmine Gravino, Gemma Catolino, and Fabio Palomba. 2025. RECOVER: Toward Requirements Generation from Stakeholders' Conversations. IEEE Transactions on Software Engineering (2025).

[38] Gianmario Voria, Giulia Sellitto, Carmine Ferrara, Francesco Abate, Andrea De Lucia, Filomena Ferrucci, Gemma Catolino, and Fabio Palomba. 2025. Fairness-aware practices from developers' perspective: A survey. Information and Software Technology 182 (2025), 107710. doi:10.1016/j.infsof.2025.107710

[39] Xin Xu, Wei Xu, Ningyu Zhang, and Julian McAuley. 2025. BiasEdit: Debiasing Stereotyped Language Models via Model Editing. In Proceedings of the 5th

Workshop on Trustworthy NLP (TrustNLP 2025). 166–184.
[40] Zeping Yu and Sophia Ananiadou. 2025. Understanding and mitigating gender bias in llms via interpretable neuron editing. arXiv preprint arXiv:2501.14457 (2025).
[41] Jerrold H Zar. 2005. Spearman rank correlation. Encyclopedia of biostatistics 7 (2005).
[42] Mengdi Zhang and Jun Sun. 2022. Adaptive fairness improvement based on causality analysis. In Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 6–17.
[43] Ting Zhang, Ivana Clairine Irsan, Ferdian Thung, and David Lo. 2025. Revisiting sentiment analysis for software engineering in the era of large language models. ACM Transactions on Software Engineering and Methodology 34, 3 (2025), 1–30.