Bias Dynamics in BabyLMs: Towards a Compute-Efficient Sandbox for Democratising Pre-Training Debiasing

Filip Trhlik, Andrew Caines, Paula Buttery
Department of Computer Science & Technology, University of Cambridge, U.K.
ALTA Institute, University of Cambridge, U.K.
ft360@cam.ac.uk

Abstract

Pre-trained language models (LMs) have, over the last few years, grown substantially in both societal adoption and training costs. This rapid growth in size has constrained progress in understanding and mitigating their biases. Since re-training LMs is prohibitively expensive, most debiasing work has focused on post-hoc or masking-based strategies, which often fail to address the underlying causes of bias. In this work, we seek to democratise pre-model debiasing research by using low-cost proxy models. Specifically, we investigate BabyLMs, compact BERT-like models trained on small and mutable corpora that can approximate bias acquisition and learning dynamics of larger models. We show that BabyLMs display closely aligned patterns of intrinsic bias formation and performance development compared to standard BERT models, despite their drastically reduced size. Furthermore, correlations between BabyLMs and BERT hold across multiple intra-model and post-model debiasing methods. Leveraging these similarities, we conduct pre-model debiasing experiments with BabyLMs, replicating prior findings and presenting new insights regarding the influence of gender imbalance and toxicity on bias formation. Our results demonstrate that BabyLMs can serve as an effective sandbox for large-scale LMs, reducing pre-training costs from over 500 GPU-hours to under 30 GPU-hours. This provides a way to democratise pre-model debiasing research and enables faster, more accessible exploration of methods for building fairer LMs.

1 Introduction

The recent dramatic increase in investment in large language models (LLMs) has driven sharp performance gains (Korinek and Vipra, 2024) while altering the way research is conducted. Since 2019, parameter counts have grown by roughly three orders of magnitude (Radford et al., 2019; Meta AI, 2025), notably inflating experimental costs and straining an academic ecosystem built on small, incremental advances (Cottier et al., 2024; Sathish et al., 2024).

This problem is particularly pressing in LM bias research, which examines how LLMs treat users differently on the basis of protected attributes – e.g., gender (Zhao et al., 2024) and ethnicity (Field et al., 2021). Despite the ubiquity of LMs, bias-removal research remains relatively linear and, constrained by the high cost of training, focuses mostly on post-hoc debiasing of already trained LMs, altering models only after their internal structures and reasoning circuits have formed (Gallegos et al., 2024). While these methods can help, they often merely mask biases detectable by our simplistic probes rather than fully removing them (Gonen and Goldberg, 2019; Gupta et al., 2024).

By contrast, debiasing strategies before or during pre-training remain rare and are outdated, partly because even a single pre-training run of a relatively simple BERT-style model exceeds 500 GPU hours on current hardware (Devlin et al., 2019). This work investigates whether significantly less costly models could replace standard LMs in debiasing research, focusing on models from the BabyLM Challenge, a research initiative that seeks to create well-performing models trained on small-scale datasets (Warstadt et al., 2023a; Charpentier et al., 2025). These properties make BabyLMs a promising democratisation tool, combining relevant performance that matches or surpasses BERT, mutable corpora that can be readily experimented with, and much lower pre-training time. In this paper, we verify this proposition, showing that BabyLMs:

1. *acquire and express biases in a way that is representative of standard LMs*
2. *respond to established debiasing methods similarly to standard LMs*
3. *enable the democratisation of pre-model debiasing research*

Findings of the Association for Computational Linguistics: ACL 2026, pages 33356–33377

2 Background

2.1 BabyLMs

BabyLMs belong to the field of low-resource language models that seeks to democratise NLP with well-performing, affordable LMs (Van Nguyen et al., 2024; Warstadt et al., 2023b).

They are inspired by the fact that human children encounter three to four orders of magnitude less linguistic data than conventional LMs, with Llama-3 using approximately 15 trillion tokens (Grattafiori et al., 2024) while a 13-year-old child may only have encountered about 100 million words (Gilkerson et al., 2017). Unlike approaches that reduce parameter counts (Hoffmann et al., 2022), the BabyLM task restricts the amount of pre-training data, simulating a more human-like learning environment.

It utilises the aforementioned 100 million words as the training corpus, which contains transcriptions of child-directed speech, child-oriented texts (e.g., books, subtitles), and other commonly available data (e.g., Wikipedia) (Warstadt et al., 2023b).

2.2 Methods for Evaluating LM Performance

When introducing alternative LM architectures, we need evaluation frameworks to compare performance. **BLiMP** probes core grammatical knowledge via minimal pairs across multiple categories (syntax, semantics, etc.) (Warstadt et al., 2020). Additionally, the **BabyLM BLiMP supplement** adds five extra dialogue/question probes (Warstadt et al., 2023a). **EWoK** evaluates cognition-inspired world-model knowledge by asking models to match two contexts to two targets, discouraging reliance on surface likelihoods (Ivanova et al., 2024). Lastly, more compute-intensive **GLUE/SuperGLUE** provide an overview of an LM's downstream capabilities (Wang et al., 2018, 2019).

2.3 Notable BabyLM Architectures

Throughout the years, numerous LM variants have been submitted to the competition, with the most relevant ones listed here. The **LTG-BERT** architecture (Samuel et al., 2023) keeps the BERT backbone but adds NormFormer (Shleifer et al., 2021), gated GELU (Shazeer, 2020), disentangled relative positions (He et al., 2020), and span masking (Joshi et al., 2020), outperforming BERT with a much smaller corpus for an identical number of training steps. **GPT-BERT** (Charpentier and Samuel, 2024), the current SOTA BabyLM, retains LTG-BERT's architecture but adds a hybrid objective that combines span masking with causal next-token prediction (BehnamGhader et al., 2024), which further separates it from the classic BERT. Other approaches have also experimented with preprocessing (Cheng et al., 2023) or curriculum learning (Diehl Martinez et al., 2023). However, they underperformed the LTG-based architectures.

2.4 Frameworks for Analysing LM Bias

Blodgett et al. (2020) define bias as systematic patterns in representations or outputs that reinforce social inequalities, resulting in allocational or representational harms. Frameworks for evaluating LM bias range from intrinsic probes to more costly, task-specific extrinsic evaluations, with only a limited correlation observed between the two (Goldfarb-Tarrant et al., 2020; Cao et al., 2022).

StereoSet measures intrinsic bias by making the model rank stereotypical, anti-stereotypical, and unrelated continuations (Nadeem et al., 2020). The bias score is derived from the ratio of examples in which the model prefers stereotypical over anti-stereotypical options, with an unbiased model yielding 50. Another notable dataset, **CrowS-Pairs** (Nangia et al., 2020), uses entire stereotype/anti-stereotype sentence pairs, with the score again reflecting the proportion of stereotypical sentences chosen. Other frameworks include SEAT (May et al., 2019), which provides a lightweight intrinsic option, and WinoBias (Zhao et al., 2018a), which targets extrinsic coreference bias.

2.5 Techniques for LM Bias Mitigation

To examine techniques for reducing biases in LMs, we follow Guo et al. (2024) and organise them by application stage. *Post-model approaches*, which leave the model intact and adjust only representations, are cheap and interpretable but are surface-level: **Iterative Nullspace Projection** (INLP) removes linearly decodable protected-attribute signals (Ravfogel et al., 2020) and **Sent-Debias** projects away PCA-estimated bias subspaces (Liang et al., 2020). Nevertheless, nonlinear probes frequently reveal residual bias, underscoring that post-hoc fixes are linear and incomplete (Sun et al., 2025; Gonen and Goldberg, 2019).

Intra-model methods involve fine-tuning that debiases the full model: increasing **dropout** can disrupt biased representations but risks negative performance impacts (Webster et al., 2020); **Counterfactual Data Substitution** (CDS) rewrites biased instances (e.g., swapping gendered entities), placing them into a new context (Bartl et al., 2020; Webster et al., 2018); and **debiasing losses** that motivate the model to debias itself (Park et al., 2023). While these methods bring some improvement, other studies have demonstrated their brittleness (Mendelson and Belinkov, 2021).

Pre-model methods modify data or pre-training to reduce bias before training even starts. They yield more stable debiasing but are costly to implement and research due to the need to retrain the model (Li et al., 2024; Xie and Lukasiewicz, 2023). Key tactics include **Counterfactual Data Augmentation** (CDA) that swaps demographic markers to rebalance the entire training corpus (Lu et al., 2018; Zmigrod et al., 2019; Webster et al., 2020); **toxic-content filtering** that displays a beneficial debiasing effect when applied (BigScience-Workshop et al., 2022; Ranaldi et al., 2024); and **perturbation augmentation**, which stochastically edits sentences along gender, ethnicity, and age axes to produce fairer models (Qian et al., 2022).

3 Experiment Setup

As established in the previous section, pre-model debiasing offers the most stable mitigation, addressing inner representations rather than surface cues. Yet these methods remain under-explored because they require costly re-training of a model from scratch and the manipulation of large, often improperly specified, corpora. Given the advantages of BabyLM architectures (competitive performance, compact datasets, and inexpensive pre-training), together with established bias- and performance-evaluation frameworks, we argue that BabyLMs can provide a practical sandbox for systematic pre-model debiasing and lower the research barrier.

To validate this proposition, we first need to understand how the debiasing behaviours of standard LMs and BabyLMs align. Therefore, we must identify candidate BabyLMs that suitably replicate the bias dynamics of standard models such as BERT.

3.1 Metrics

We utilise the performance and bias scores from frameworks described in Sections 2.2 and 2.4. For bias, this means using CrowS-Pairs and StereoSet. Both share a mathematically similar approach and a score scale. Since BabyLMs can be either masked or continuation LMs and lack next-sentence prediction, we exclude StereoSet's inter-sentence portion (Ranaldi et al., 2024).

A bias evaluation is not sufficient on its own. Since there is an established relationship between the model's biases and its performance (Nadeem et al., 2020), we estimate the performance through BLiMP, BabyLM BLiMP supplement, and EWoK.

Thus, we obtain three performance metrics and two bias metrics, all of which capture only part of a model's behaviour. They all share the same scale but probe the model with different sentences and contexts. Therefore, they reveal different parts of its bias and performance profile (Zakizadeh and Pilehvar, 2025). To obtain a more comprehensive picture, we average the individual performance scores into a *composite performance* metric and the two bias scores into a *composite bias* metric (metrics composition details in Appendix B).

3.2 Candidate BabyLM

As noted, our aim is to select a BabyLM that comes closest behaviour-wise to standard LMs, while still retaining desirable characteristics, such as low-cost training. Firstly, this requires showing that BabyLMs in general display bias acquisition dynamics; specifically, that they acquire more biases with increased performance, as observed within larger LMs (Nadeem et al., 2020).

We do this by evaluating composite bias and performance metrics for every notable BabyLM and various variants of standard LMs, with all models used listed in Appendix A. This yields Table 1, which shows the correlation between *composite performance* and *composite bias* for both model classes. The strong positive correlation, and its shared strength across BabyLMs and standard LMs, shows that the overall trend of bias increasing with performance is preserved in BabyLMs, with the exact spread of models shown in the Appendix C. Therefore, BabyLMs can be a valid sandbox for studying debiasing techniques with resources that are feasible for many research groups.

Model Class	N	r(**Composite Performance, Composite Bias**)
BabyLM	9	0.833
Standard	16	0.753

Table 1: Pearson correlation between composite performance and bias for the different model classes

With BabyLM eligibility established, we select candidate models. The SOTA variant of the LTG-

BERT architecture, `ltg-bert-babylm`, is closest to the original BERT. However, it has been trained for 1,500 epochs, requiring roughly the same GPU-hours as the original BERT model, making it unsuitable for our purposes. Thus, we also select its low-resource variant, `ltgbert-100m-2024`, trained for just 40 GPU-hours. Although more distant from BERT, it still achieves reasonable performance and exhibits measurable bias, making it a promising low-cost BabyLM. In conclusion, the two candidate models for further study are **LTG-BERT** (ltg-bert-babylm) and **LTG-Baseline** (ltgbert-100m-2024). LTG-BERT tests whether any BabyLM can sufficiently replicate standard debiasing patterns, while LTG-Baseline tests whether these hold even with limited training.

4 Model Viability

Having established that BabyLMs acquire biases in a similar way to standard LMs, the next step is to test whether they also debias comparably. We analyse whether their pre-training corpora are comparable in terms of the biases acquirable from them. We also investigate how their composite performance and bias change in reaction to a wide selection of intra-model and post-model debiasing techniques, each targeting different parts of the models.

4.1 Corpora

Debiasing strategies, especially pre-model ones, largely entail altering the pre-training corpus; thus, we must demonstrate that the BabyLM corpus can support the emergence of the same bias dynamics observed in BERT. We therefore examine corpus aspects known to induce LM biases. This section reports the most critical results; the remainder is listed in Appendix D.

In terms of the corpora, we utilise the 2023 BabyLM challenge corpus, used to train LTG-BERT, and closely replicate the unavailable BERT corpus from a Wikipedia dump (Broad, 2022) and a BookCorpusOpen re-crawl (Di Liello, 2022; Bandy and Vincent, 2021), preserving the two-corpora token ratio and the original ~3B-token size.

With the corpora established, we examine differences in their topical coverage (Chang et al., 2019; Zhao et al., 2019). For each topic (e.g., race, gender), we compute the percentage of each corpus formed by keywords related to its subcategories (e.g., male, female) (Meade et al., 2021). Gender is the most prominent category. Table 2 shows that the male gender is more represented than the female gender across corpora, with the resulting models also biased in a male-centric direction. This suggests that they should react to gender-focused debiasing in the same way.

Category	BabyLM	BERT
Male	2.07%	1.83%
Female	1.03%	1.30%

Table 2: Gender representation

Across other bias categories, most trends are shared between the corpora. In ethnicity-term frequency, Caucasian terms are consistently most over-represented, Black-related terms second, and Asian strongly last. In religion, Christian terms dominate, while Jewish terms are barely represented in both. LGBTQ+ topics appear at similar frequencies. The differences lie in the BERT corpus representing Black and Muslim terms much more substantially. Overall, despite BERT containing higher frequencies of biased terms across categories, the BabyLM corpus largely preserves similar topic ratios, indicating it ought to support debiasing techniques.

Category	BabyLM	BERT
Black	0.035%	0.062%
Caucasian	0.049%	0.067%
Asian	0.014%	0.034%
Jewish	0.008%	0.007%
Christian	0.034%	0.065%
Muslim	0.007%	0.028%
LGBTQ+	0.017%	0.022%

Table 3: Demographic/religion mentions

Finally, as toxicity and hate-speech are linked to increased biases (BigScience-Workshop et al., 2022), we used established models (Hanu and Unitary AI, 2020; Antypas and Camacho-Collados, 2023) to label their presence in each corpus. Table 4 shows that the BabyLM corpus is more toxic and hateful, even containing blatantly racist and sexualised terms, contrary to its child-aligned disposition.

Corpus	Toxic	Hate-speech
BabyLM	3.12%	0.55%
BERT	0.95%	0.34%

Table 4: Toxicity and hate-speech sentence rates

Overall, compared to the BERT corpus, BabyLM exhibits similar topic frequencies and a higher presence of toxic and hateful sentences, enabling the same bias and debiasing dynamics. Consequently, the BabyLM corpus appears capable of supporting both bias acquisition and debiasing similar to those observed with BERT.

4.2 Debiasing behaviour

Building on the finding that the BabyLM corpus aligns with the BERT corpus across all bias-relevant metrics, we test whether the two model classes debias comparably. To analyse behavioural overlaps, we use a broad set of intra- and post-model debiasing techniques, each using different mechanisms and targeting distinct model components. In each test, we compare how debiasing shifts LTG-BERT's and LTG-Baseline's composite bias and performance relative to BERT.

Starting with post-model debiasing, we apply two debiasing methods, **Sent-Debias** and **INLP**, both targeting trained models with already frozen encoders. Sent-Debias learns a gender subspace from text and subtracts it from the final hidden representations. INLP trains linear classifiers for protected attributes and iteratively projects out the directions that make those attributes linearly separable. The implementation of both approaches uses a 2.5M-word Wikipedia dump for their debiasing signals (Meade et al., 2021).

Looking at the impact of debiasing in Figure 1, we see that Sent-Debias consistently reduces composite gender bias across all models, although the performance impact varies. INLP remains consistent across architectures regarding its effects on bias. Gender-focused INLP lowers overall bias, while race-focused INLP does not reduce composite bias and also harms accuracy. The gender-focused INLP's performance effects align with model fit: the over-fitted LTG-BERT benefits from it as a form of regularisation, achieving SOTA performance (Appendix E), the more under-fitted LTG-Baseline loses useful information, and full-data BERT incurs no severe penalties. With this behavioural nuance, we conclude that the methods' impact on bias is consistent across models.

Next, we probe intra-model debiasing with four strategies. **CDA** balances a 10M-word Wikipedia dump by duplicating sentences containing gendered or racial terms and swapping them (Meade et al., 2021). **CDS** uses the gender-balanced corpus, substituting each gender mention with the opposite to create anti-stereotype contexts (Webster et al., 2018; Bartl et al., 2020). A **debiasing-loss** setup

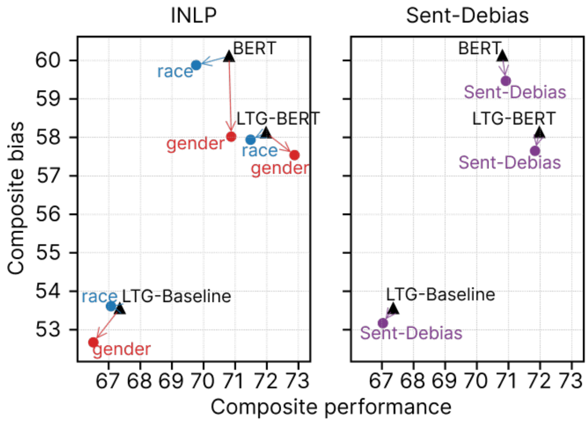

Figure 1: Bias and performance changes caused by post-model debiasing strategies

trains on the Gender Pronoun Resolution task with Stereotype Neutralisation and Elastic Weight Consolidation (Zhao et al., 2018b; Park et al., 2023). A **dropout** variant increases dropout during continued pre-training on the same Wikipedia dataset.

Across methods, Figure 2 shows that the models shift in the same direction, differing mainly in magnitude. CDA on gender and race reduces composite bias across models, with stronger gender effects and similar performance-loss trends. CDS again yields near-identical bias reduction across models with modest, method-consistent accuracy costs. Debiasing loss induces the largest accuracy drop and bias decrease, with LTG-Baseline tracking BERT most closely. Dropout yields weaker debiasing while preserving the trends; some noise in the performance–bias-loss ratio is expected since BERT lacks the extra normalisation of LTG models (Shleifer et al., 2021), leading to over-confident logits that drift under perturbation (Gal and Ghahramani, 2016; Kong et al., 2020).

These convergent effects show that BabyLMs simulate the behaviour of post- and intra-model debiasing techniques on BERT. We quantify alignment by pairing each model's performance and bias shifts per method and running canonical correlation analysis (Hotelling, 1936). The results in Table 5 show that LTG-Baseline is the most faithful proxy of BERT's debiasing behaviour, whereas LTG-BERT's over-fitting likely dampens alignment.

5 Pre-Model Debiasing Experiments

In the previous section, we showed that standard LMs and BabyLMs share debiasing dynamics, enabling us to estimate a debiasing method's impact on a standard LM by applying it to a BabyLM. We can now utilise this to run pre-model debias-

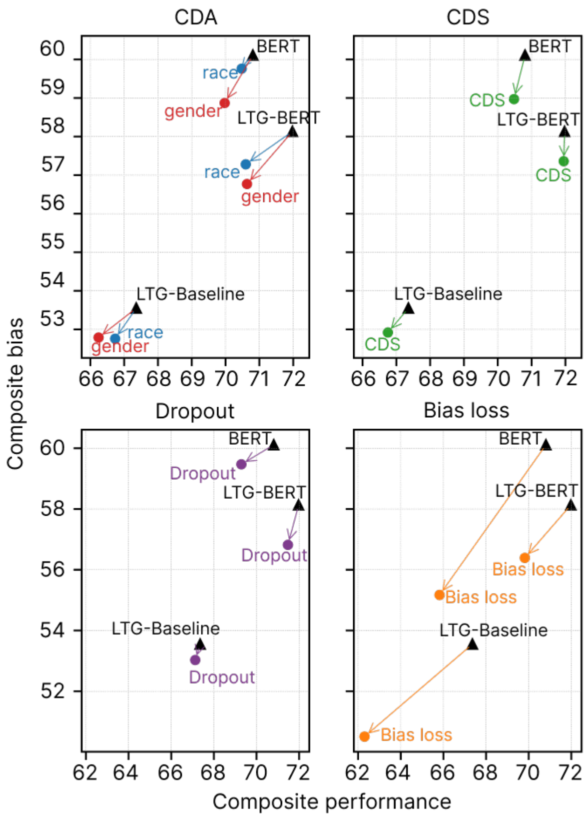

Figure 2: Bias and performance changes caused by intra-model debiasing strategies

Model pair	Correlation(ρ_1)
BERT↔LTG-BERT	0.796
BERT↔LTG-Baseline	**0.981**
LTG-BERT↔LTG-Baseline	0.772

Table 5: First canonical correlation (ρ_1) between performance and bias shifts across all debiasing methods

ing experiments on LTG-Baseline instead of BERT, reducing the cost from 500 GPU-hours per experiment to just over 30.

Throughout this section, we reinforce that BabyLM mimics BERT's reported behaviour and show how this benefits future pre-model debiasing investigations. All experiments use the LTG-Baseline architecture and alter only the training corpus (pre-training setup details in the Appendix F).

We establish a baseline by training the LTG model on the original BabyLM corpus, tracking performance and bias over time. During this baseline training, the model picked up bias early, which then stabilised at a steady level, while performance improved more gradually as it learned richer linguistic structure. The quick uptake of bias suggests that the biases likely stem from the topical imbalances and stereotypes, which most pre-model debiasing techniques target.

5.1 CDA Pre-model Debiasing

As a first experiment, we apply pre-model CDA: for every sentence containing a gendered term, we append a flipped-gender counterpart, increasing corpus size by ∼59% and creating a gender-balanced corpus. To check whether any pre-training effects come from balancing rather than simple duplication, we run an ablation that duplicates an equal number of random BabyLM corpus sentences. Figure 3 shows that the CDA initially slows the model's grasp of some linguistic concepts but, with longer training, reaches near-baseline performance while clearly curbing bias by preventing its observed steady growth. In contrast, the ablation results in the same performance drop while producing only a small bias reduction. Overall, this validates CDA as a sound debiasing strategy and highlights BabyLMs as a cost-effective platform for such controlled experiments.

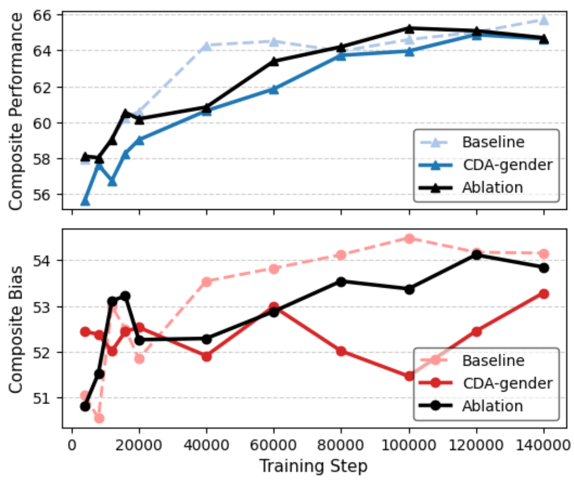

Figure 3: Bias and performance metrics evaluation from pre-training on the CDA and CDA-ablation corpora

To assess the robustness of the BabyLM-based sandbox, we repeat CDA pre-training with multiple random seeds while keeping everything fixed. Figure 4 shows consistent bias and performance trajectories across runs, with only minor variation and no change to the overall conclusions.

5.2 Toxicity Removal

Next, we examine the suggested but unproven claim that corpus toxicity directly drives model bias within LLMs. With 3.39% of BabyLM sentences being toxic or hateful, we seek to push toxicity to 0% via two interventions.

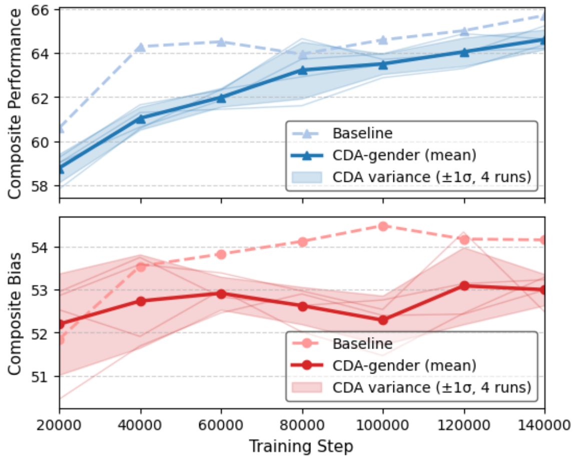

Figure 4: Bias and performance variance during CDA pre-training across random seeds {1, 10, 50, 100}

Firstly, we utilise an LLM (Llama-3.3-70B) to rewrite toxic sentences while preserving text meaning, discarding only 0.42% of sentences that could not be detoxified (implementation detailed in Appendix G). This yielded a slight performance gain and a surprisingly small bias drop, possibly because it imported the LLM's style and biases into the corpus. Secondly, we dropped all toxic sentences, which significantly reduced bias and slightly harmed performance.

Lastly, our ablation test removing an equal number of non-toxic sentences matched the performance drop but failed to match the bias decrease, showing that eliminating toxicity itself, rather than corpus shrinkage, drives the debiasing. Figure 5 summarises these results, establishing a clear link between toxicity and standard bias and highlighting the advantage of our BabyLM-based approach that allows us to identify such behaviour.

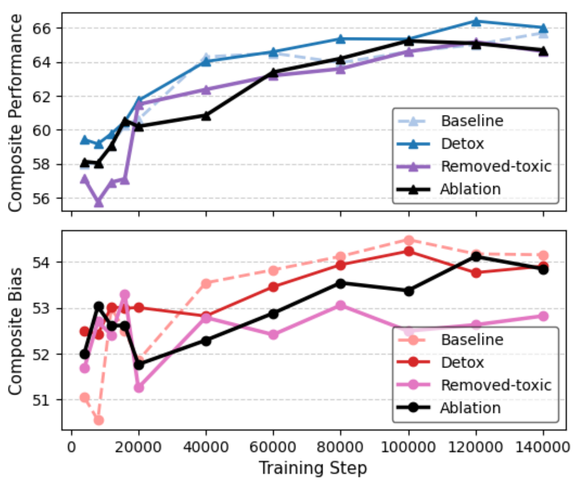

Figure 5: Evolution of bias and performance during pre-training under three corpus strategies: LLM detoxification, removing toxicity, and toxicity-removal ablation

5.3 Perturbation Augmentation

Finally, we evaluate perturbation augmentation (Qian et al., 2022), which uses a perturber LM to rewrite the corpus by randomly swapping demographic references (race and gender), thereby equalising topic–demographic co-occurrence (implementation discussed in Appendix H).

In the original study, pre-training RoBERTa on perturbed data reduced bias substantially with a slight performance gain. Using ∼800× fewer GPU-hours, we closely reproduce these trends (Figure 6). The perturbation outperforms CDA in debiasing, likely by introducing greater lexical and syntactic variety and by covering more attributes. Moreover, it boosts performance, with the small gains possibly reflecting improved word order in lower-quality sentences caused by perturbation. Overall, even with this more complex debiasing strategy, BabyLM closely mirrors behaviour reported in far more resource-intensive experiments.

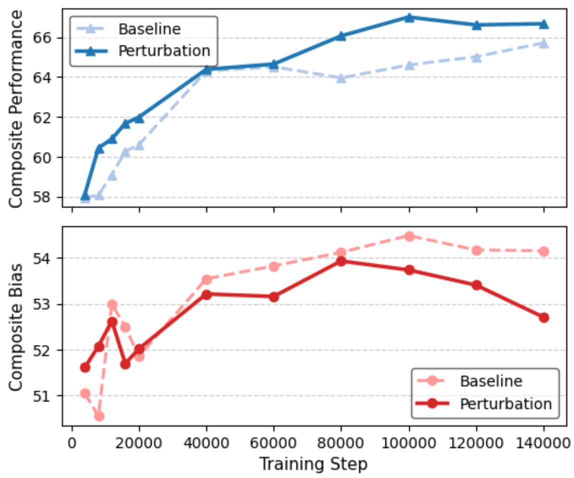

Figure 6: Bias and performance metrics evolution from pre-training on the perturbed corpus

In conclusion, LTG-Baseline successfully reproduced the expected behaviours across all tested pre-model debiasing tasks. Moreover, this setup let us both replicate the original experiments and introduce ablations that examined key components of several debiasing strategies. Consequently, under a strict budget, we were able to conduct experiments that had not previously been attempted.

Thus, we show that BabyLMs, and possibly other similarly positioned models, can be very versatile and representative tools for ascertaining the effectiveness of debiasing methods, greatly lowering both the cost and the time required to conduct the exploration stage of debiasing experiments.

6 Findings

With our analysis complete, we note several key trends that directly support the use of low-cost LMs to democratise research on pre-model debiasing.

Finding 1: BabyLM corpora and architectures share the same tendencies regarding bias–performance dynamics as standard LMs. Despite far smaller training sets and adapted architectures, BabyLMs acquire linguistic knowledge and biases along the same trajectories as standard BERT-style LMs. Across models, *composite performance* and *composite bias* are strongly and similarly correlated (Table 1). Corpus analysis shows that the BabyLM data, although grammatically simpler, contains the same well-known bias instigators (gender imbalance, topical skew, toxicity, etc.). In particular, male terms are over-represented, Christian terms dominate religious mentions, and toxicity and hate-speech rates are higher than in our reconstructed BERT corpus. Consequently, the SOTA BabyLMs match both BERT's performance and bias, while lower-resource BabyLMs preserve the same bias–ability correlation. This makes such models viable low-cost sandboxes for bias studies.

Finding 2: The debiasing behaviour of BabyLMs strongly correlates with that of standard LMs. Building on the first finding, we compared post-model and intra-model debiasing across BERT, LTG-BERT, and LTG-Baseline to test whether BabyLMs share the same debiasing behaviour as standard LMs. Across all debiasing methods, BabyLMs exhibited decreases in bias closely resembling BERT's. Greater variance, on the other hand, was observed in performance changes, with post-model approaches being especially sensitive to architectural and pre-training differences between models. Nevertheless, canonical correlation analysis of bias and performance shifts across all methods showed near-perfect alignment between BERT and LTG-Baseline and strong alignment with LTG-BERT. Thus, we show that even a BabyLM trained for 30 GPU-hours closely proxies debiasing dynamics observed in full-scale models.

Finding 3: BabyLMs enable informative and far more cost-effective pre-model debiasing research. Finally, we demonstrate the utility of our proposed approach by running seven pre-training interventions on LTG-Baseline, reproducing established results while enabling new experiments and targeted ablations. We use these interventions both to validate our method and to illustrate how it can be used to benefit pre-model debiasing research. We replicate prior findings that CDA reduces bias but can slow learning, and that perturbation augmentation yields larger bias reductions without degrading performance. Beyond replication, we directly linked corpus toxicity to downstream bias, the first study in a pre-trained model setting to directly test, rather than merely imply, the effect of removing toxic sentences on the resulting bias. With additional ablation experiments, we then isolated specific bias causes (gender imbalance, toxicity), showing that debiasing helps primarily by addressing these instigators rather than incidental corpus restructuring. Crucially, the same bias–performance trends persist across runs with different random seeds, indicating that the observed effects are not a one-off consequence of training stochasticity and supporting BabyLMs as a stable, cost-effective sandbox for pre-model debiasing research under limited compute.

7 Conclusion

This work raised the issue that most LM debiasing research focuses only on pre-trained models with already-formed biased circuits. We argued that, to develop effective debiasing strategies, we must first understand how bias emerges in LMs and devise approaches that prevent its original formation. Such research is rare due to its prohibitive cost. To fix this, we proposed investigating debiasing dynamics in well-performing, low-resource, data-efficient models, such as BabyLMs.

Our experiments showed that BabyLMs use sufficient data and acquire intrinsic biases comparably to standard LMs when matched for performance. Their debiasing behaviour likewise mirrors that of other LMs, indicating that BabyLMs' democratised pre-training setup does not disqualify them. Thus, we moved from pre-training costs of 500 GPU-hours to 30 GPU-hours using a BabyLM, creating a pathway for affordable debiasing research. With this, we reproduced previously reported results and added findings that solidified toxicity and bias imbalance as major contributing factors to LM bias.

Overall, our hope is that this research encourages the use of low-cost LMs that enable the exploration and negation of bias formation, allowing researchers to identify promising methods before committing to costly large-scale experiments and to move the entire LM debiasing field forward.

8 Limitations

One limitation of this study is the set of metrics it was able to use. BabyLMs limit us to more simplistic bias and performance-evaluation frameworks, since they do not offer sufficiently developed language and world understanding to support advanced extrinsic evaluations. Furthermore, due to the large amount of evaluation throughout the entire paper, frameworks like SuperGLUE, which take 2–3 hours to run on our hardware, are infeasible. In addition, all the metrics we used are English-specific. This is important to note because, even though non-English versions of these benchmarks exist (Névéol et al., 2022; Öztürk et al., 2023), they are very limited and still do not cover low-resource languages, which is a major obstacle for LM bias research.

Secondly, it is well discussed that bias evaluations represent only a subset of the larger issue, and there might be biases where the different models' behaviours actively differ but cannot be observed. Thus, we recommend using the composite bias metric to identify overall trends, but there may be a need to investigate or propose more specialised tests when tracing specific types of biases.

Related to this, while BabyLMs show promising alignment, there will be tasks that they cannot replicate. As noted in the paper, we propose them as a tool for *exploration*, helping to identify promising debiasing strategies. When these strategies are identified, they still need to be validated on larger models. Nevertheless, BabyLMs still allow us to skip the costly experimental phase.

Finally, it should be noted that BabyLMs were used as a promising and easily available set of architectures that displayed appropriate properties for the task at hand. We still encourage efforts to explore different corpora and architectures. CLMs, especially in the form of LLMs, cannot be replaced by MLMs. As such, future work should propose a model that democratises debiasing research for CLMs. However, MLMs remain widely deployed in classification- and retrieval-heavy industry pipelines (Clavié et al., 2025), and with only two pre-training debiasing strategies ever proposed, their debiasing remains a relevant and underexplored avenue of research. Likewise, there might be corpora even better suited for testing debiasing methods than the BabyLM one. This study simply shows that the promising results obtained with the LTG-Baseline and the BabyLM corpus provide strong evidence that this path towards the democratisation of debiasing research is possible and promising.

9 Ethical Considerations

This work does not involve human-subject data or sensitive personal information. Nevertheless, any discussion of bias and toxicity in language models must acknowledge that our evaluation is incomplete and grounded in prior definitions of bias. Accordingly, we do not claim that any debiasing approach is definitive, nor that even the best-performing strategies remove all forms of bias.

As noted, implementations of our framework should not proceed under the assumption that it is sufficient on its own. In particular, when debiasing a publicly available model, we must ensure that biases are also evaluated in the model itself rather than purely relying on cheaper proxies.

Acknowledgements

This paper reports on work supported by Cambridge University Press & Assessment. We thank colleagues at the ALTA Institute, as well as Jeremy Lo Ying Ping, Patrick Haller, and Simon Sorg, for their support and feedback.

References

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242, Toronto, Canada. Association for Computational Linguistics.

Jack Bandy and Nicholas Vincent. 2021. Addressing" documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. *arXiv preprint arXiv:2105.05241*.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and

Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

BigScience-Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Nicholas Broad. 2022. English wikipedia dump: 2022-03-01 (wiki-20220301-en). Kaggle dataset. Accessed 11 Aug 2025.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. 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): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, and 1 others. 2025. Babylm turns 3: Call for papers for the 2025 babylm workshop. *arXiv preprint arXiv:2502.10645*.

Lucas Georges Gabriel Charpentier and David Samuel. 2024. Gpt or bert: why not both? *arXiv preprint arXiv:2410.24159*.

Ziling Cheng, Rahul Aralikatte, Ian Porada, Cesare Spinoso-Di Piano, and Jackie CK Cheung. 2023. McGill BabyLM shared task submission: The effects of data formatting and structural biases. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 207–220, Singapore. Association for Computational Linguistics.

Benjamin Clavié, Nathan Cooper, and Benjamin Warner. 2025. It is all in the [mask]: Simple instruction-tuning enables bert-like masked language models as generative classifiers. *Natural Language Processing Journal*, 11:100150.

Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, Tamay Besiroglu, and David Owen. 2024. The rising costs of training frontier ai models. *arXiv preprint arXiv:2405.21015*.

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)*, pages 4171–4186.

Luca Di Liello. 2022. Bookcorpusopen. Hugging Face dataset. Revision: main@edb74e6. Accessed 11 Aug 2025.

Richard Diehl Martinez, Zébulon Goriely, Hope McGovern, Christopher Davis, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2023. CLIMB – curriculum learning for infant-inspired model building. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 112–127, Singapore. Association for Computational Linguistics.

Paul Ekman and 1 others. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. 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)*, pages 1905–1925, Online. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

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):1097–1179.

J. Gilkerson, J. A. Richards, S. F. Warren, J. K. Montgomery, C. R. Greenwood, D. Kimbrough Oller, J. H. L. Hansen, and T. D. Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, 26(2):248–265.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2020. Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915*.

Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca Passonneau. 2024. Sociodemographic bias in language models: A survey and forward path. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 295–322, Bangkok, Thailand. Association for Computational Linguistics.

Lasse Hansen, Ludvig Renbo Olsen, and Kenneth Enevoldsen. 2023. Textdescriptives: A python package for calculating a large variety of metrics from text. *Journal of Open Source Software*, 8(84):5153.

Laura Hanu and Unitary AI. 2020. Detoxify.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024a. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.

Michael Y Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024b. Findings of the second babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. *arXiv preprint arXiv:2412.05149*.

Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, RT Pramod, Gabriel Grand, and 1 others. 2024. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint arXiv:2405.09605*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Abdullatif Köksal, Omer Yalcin, Ahmet Akbiyik, M. Kilavuz, Anna Korhonen, and Hinrich Schuetze. 2023. Language-agnostic bias detection in language models with bias probing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12735–12747, Singapore. Association for Computational Linguistics.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1326–1340, Online. Association for Computational Linguistics.

Anton Korinek and Jai Vipra. 2024. Concentrating intelligence: scaling and market structure in artificial intelligence*. *Economic Policy*, 40(121):225–256.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2024. Data-centric explainable debiasing for improving fairness in pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3773–3786, Bangkok, Thailand. Association for Computational Linguistics.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, pages 5502–5515, Online. Association for Computational Linguistics.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. *arXiv preprint arXiv:1807.11714*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2021. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. *arXiv preprint arXiv:2110.08527*.

Michael Mendelson and Yonatan Belinkov. 2021. Debiasing methods in natural language understanding make bias more accessible. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1545–1557, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`. Accessed 29 Jul 2025.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Ibrahim Tolga Öztürk, Rostislav Nedelchev, Christian Heumann, Esteban Garces Arias, Marius Roger, Bernd Bischl, and Matthias Aßenmacher. 2023. How different is stereotypical bias across languages? In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 209–229. Springer.

SunYoung Park, Kyuri Choi, Haeun Yu, and Youngjoong Ko. 2023. Never too late to learn: Regularizing gender bias in coreference resolution. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 15–23, New York, NY, USA. Association for Computing Machinery.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI technical report.

Leonardo Ranaldi, Elena Sofia Ruzzetti, Davide Venditti, Dario Onorati, and Fabio Massimo Zanzotto. 2024. A trip towards fairness: Bias and de-biasing in large language models. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 372–384, Mexico City, Mexico. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: Bert meets british national corpus. *arXiv preprint arXiv:2303.09859*.

Vishwas Sathish, Hannah Lin, Aditya K Kamath, and Anish Nyayachavadi. 2024. Llempower: Understanding disparities in the control and access of large language models. *arXiv preprint arXiv:2404.09356*.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Sam Shleifer, Jason Weston, and Myle Ott. 2021. Normformer: Improved transformer pretraining with extra normalization. *arXiv preprint arXiv:2110.09456*.

Lihao Sun, Chengzhi Mao, Valentin Hofmann, and Xuechunzi Bai. 2025. Aligned but blind: Alignment increases implicit bias by reducing awareness of race. *arXiv preprint arXiv:2506.00253*.

Chien Van Nguyen, Xuan Shen, Ryan Aponte, Yu Xia, Samyadeep Basu, Zhengmian Hu, Jian Chen, Mihir Parmar, Sasidhar Kunapuli, Joe Barrow, and 1 others. 2024. A survey of small language models. *arXiv preprint arXiv:2410.20011*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023a. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors. 2023b. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Singapore.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhongbin Xie and Thomas Lukasiewicz. 2023. An empirical analysis of parameter-efficient methods for debiasing pre-trained language models. *arXiv preprint arXiv:2306.04067*.

Shuzhou Yuan, Ercong Nie, Lukas Kouba, Ashish Yashwanth Kangen, Helmut Schmid, Hinrich Schütze, and Michael Färber. 2025. Llm in the loop: Creating the paradehate dataset for hate speech detoxification. *arXiv preprint arXiv:2506.01484*.

Mahdi Zakizadeh and Mohammad Taher Pilehvar. 2025. Blind men and the elephant: Diverse perspectives on gender stereotypes in benchmark datasets. *arXiv preprint arXiv:2501.01168*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018b. 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)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang, and Zifan Qian. 2024. Gender bias in large language models across multiple languages. *arXiv preprint arXiv:2403.00277*.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Evaluated Models

In order to run our experiment, we must analyse a sufficiently large set of BabyLMs and standard LMs, so that we can observe meaningful trends.

Regarding BabyLMs, we collect every notable BabyLM that is available online with executable code, utilising our overview in Section 2.3 (Samuel et al., 2023). It should be noted that some models require altered data input pipelines, making them impractical. Likewise, several promising papers never released their models. To ensure comparability, we take only models from the `strict` track. The final list is shown in Table 6.

For standard LMs, we also need to observe how bias scales with performance. Testing just a few primary models would not establish a trend. To resolve this, we examine the popular architectures used to create BabyLMs together with other variants of these architectures trained on different corpora (e.g. legal documents (Chalkidis et al., 2020), Twitter (Barbieri et al., 2020), scientific papers (Beltagy et al., 2019)). All these models are listed in Table 7.

B Model Alignment

Running all models through our pipelines, we obtain Pearson correlations between the bias metrics and the performance metrics. Figures 7 and 8 show the correlation results for BabyLMs and standard LMs respectively. In both groups, performance and bias metrics are strongly positively correlated, while the two bias scores have a positive but more limited correlation with each other. BabyLMs show especially noisy behaviour since some of them have insufficient world knowledge to display biases. This leads us to establish the composite metrics, which more completely capture the overall behaviour.

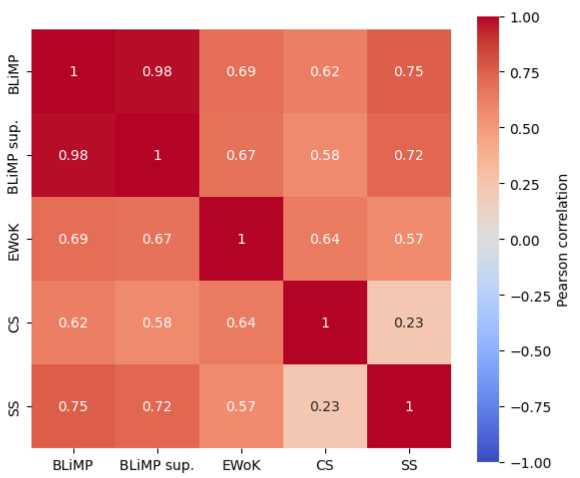

Figure 7: Pearson correlations (BabyLM models, $N = 9$)

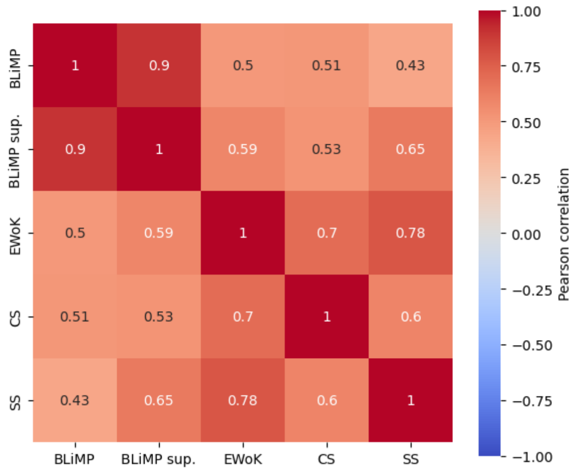

Figure 8: Pearson correlations (standard models, $N = 16$)

C Selecting Candidate BabyLMs

Furthermore, we need to select the most informative and relevant BabyLMs, which we continue using in our research. To do so, we are interested in models that are as close to the original BERT as possible in both behaviour and architecture, making them suitable for replicating BERT's behaviour. To this end, we examine Figure 9, which visualises the results obtained in the previous section, and we highlight two promising models.

Of all the models considered, ltg-bert-babylm is the most closely aligned with BERT's architecture, as indicated by the description in Section 2.3 (Samuel et al., 2023). It has no modifications that would make it more complicated or less predictable to work with.

Nevertheless, our ultimate goal is to advance democratisation research, and ltg-bert-babylm, together with the other high-performing models, was pre-trained for roughly the same number of GPU-hours as the original BERT. Thus, although ltg-bert-babylm is ideal for studying how debiasing operates in the BabyLM-BERT relation, it is less suitable for democratisation research.

This limitation was recognised even by the 2024 BabyLM organisers, who released ltgbert-100m-2024 (Hu et al., 2024a). This model is a baseline version of the SOTA ltg-bert-babylm. It was trained for only 20 epochs (versus 1,500 epochs for the SOTA model) and required less than 40 GPU-hours. Figure 10 shows the contrast between the two versions. Although ltgbert-100m-2024 is notably further from BERT in performance, it still achieves reasonable performance scores and ex-

Model key	Hugging Face ID
BabyLM2024	jdebene/BabyLM2024
elc-bert	lgcharpe/ELC_BERT_baby_100M
cambridge-climb	cambridge-climb/baseline-roberta_pre_layer_norm-model
gpt-bert-babylm	ltg/gpt-bert-babylm-base
ltg-bert-babylm	ltg/ltg-bert-babylm
ltgbert-100m-2024	babylm/ltgbert-100m-2024
roberta-base-strict-2023	babylm/roberta-base-strict-2023
babyllama-100m-2024	babylm/babyllama-100m-2024
baby-llama-2-345m	JLTastet/baby-llama-2-345m

Table 6: Examined BabyLMs with their Hugging Face IDs

Model key	Hugging Face ID
BiomedBERT-abstract	microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract
BiomedBERT-abstract-fulltext	microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext
DialoGPT-large	microsoft/DialoGPT-large
DialoGPT-medium	microsoft/DialoGPT-medium
DialoGPT-small	microsoft/DialoGPT-small
bert-base-cased	bert-base-cased
bert-base-uncased	bert-base-uncased
bert-for-patents	anferico/bert-for-patents
BNC bert	ltg/ltg-bert-bnc
finbert	yiyanghkust/finbert-pretrain
gpt2-medium	openai-community/gpt2-medium
gpt2-xl	openai-community/gpt2-xl
legal-bert	nlpaueb/legal-bert-base-uncased
roberta-base	FacebookAI/roberta-base
scibert	allenai/scibert_scivocab_uncased
twitter-roberta-base	cardiffnlp/twitter-roberta-base

Table 7: Examined standard models with their Hugging Face IDs

hibits measurable bias, making it a good baseline for our experiments.

D Corpora

This section details all other experiments conducted with the aim of exploring and comparing the corpora of BabyLMs and standard LMs.

D.1 Structural and Syntactic Metrics

We start the analysis by examining the basic linguistic metrics present in the corpora, trying to understand the composition and complexity of each dataset. This might help us understand whether the BabyLM corpus is sufficiently coherent to form a full range of biases.

To do this, we leverage the TextDescriptives library by Hansen et al. (2023), which allows us to derive more than sixty syntactic and discourse-level statistics. Because all corpora exceed the toolkit's token limit for text processing, we split the text into chunks of 8192 tokens, retaining large enough samples to assess the more structural metrics while keeping the problem tractable. In the end, the outputs per chunk are averaged into final values representing the entire corpus. This ensures that no corpus is penalised for its size while capturing the structural and syntactic profile needed for analysis.

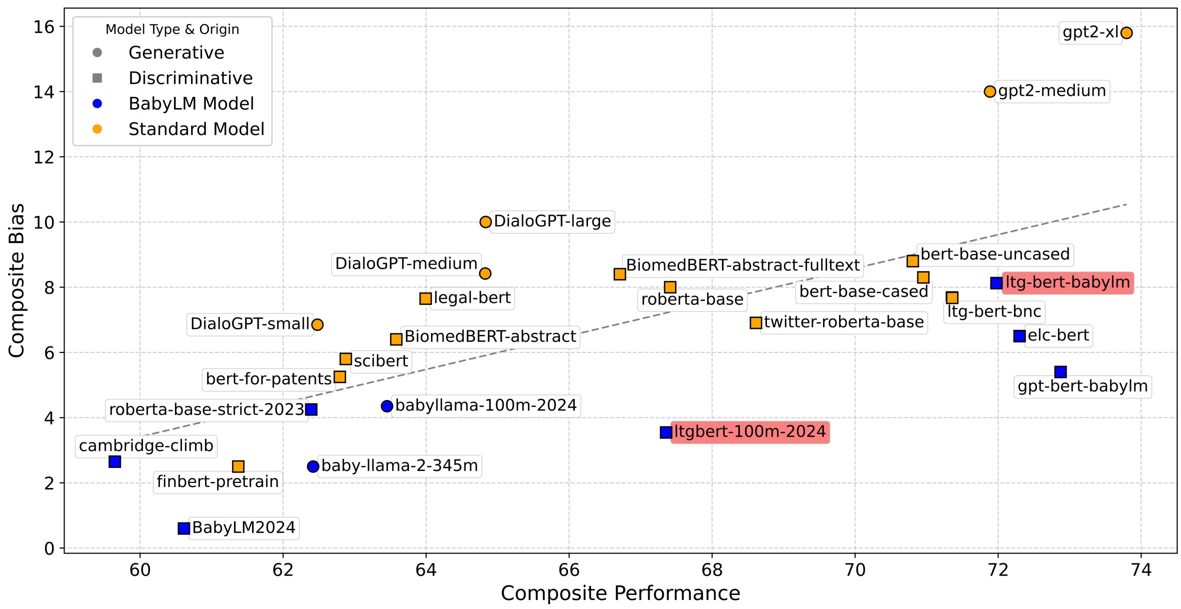

Figure 9: Average performance correlated with average bias for all evaluated LMs, with the average trend being shown and the candidate models being highlighted

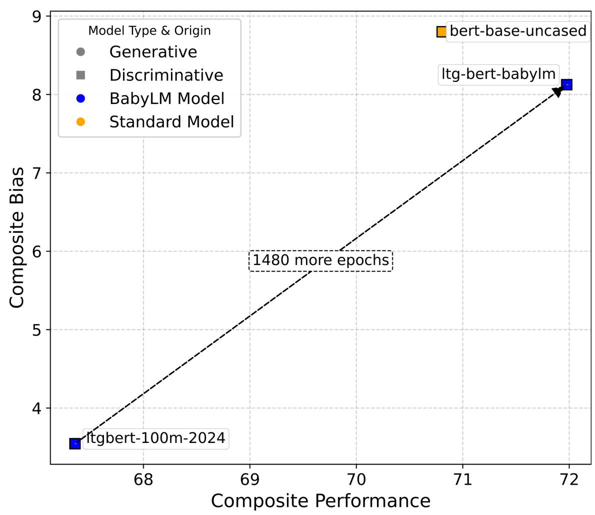

Figure 10: Performance–bias trade-off for ltg-bert-babylm and its low-cost baseline

D.1.1 Part-of-Speech Profiles

PoS statistics give us a compact high-level overview, informing us about the average type of the texts present in each corpus. We are especially interested in the frequency of function words (e.g., pronouns, determiners), which organise discourse, versus the frequency of content words (e.g., nouns, verbs), which carry meaning.

Figure 11 shows us that BabyLM displays a lower frequency of nouns and pronouns, signalling a simpler and more narrative-driven corpus.

Figure 12 serves to further highlight this contrast. BabyLM's proportion of pronouns to nouns is, unlike that of BERT, almost at parity. Thus, in the BabyLM corpus, more tokens are used to refer to individuals rather than to describe unique objects or concepts, again indicating that it is linguistically simpler than the BERT corpus.

D.1.2 Linguistic Complexity and Readability

Since it appears that the main difference is text complexity, we investigate the readability metrics.

We choose to use the Flesch–Kincaid Grade Level (FKGL), which is computed from the average sentence length and the average number of syllables per word. Its final value then corresponds to grades in the US school system, ranging between 0 and 18. With the results shown in Table 8, we can once again see that BERT has the linguistically more advanced texts and BabyLM has the simpler ones. The FKGL places the BabyLM cor-

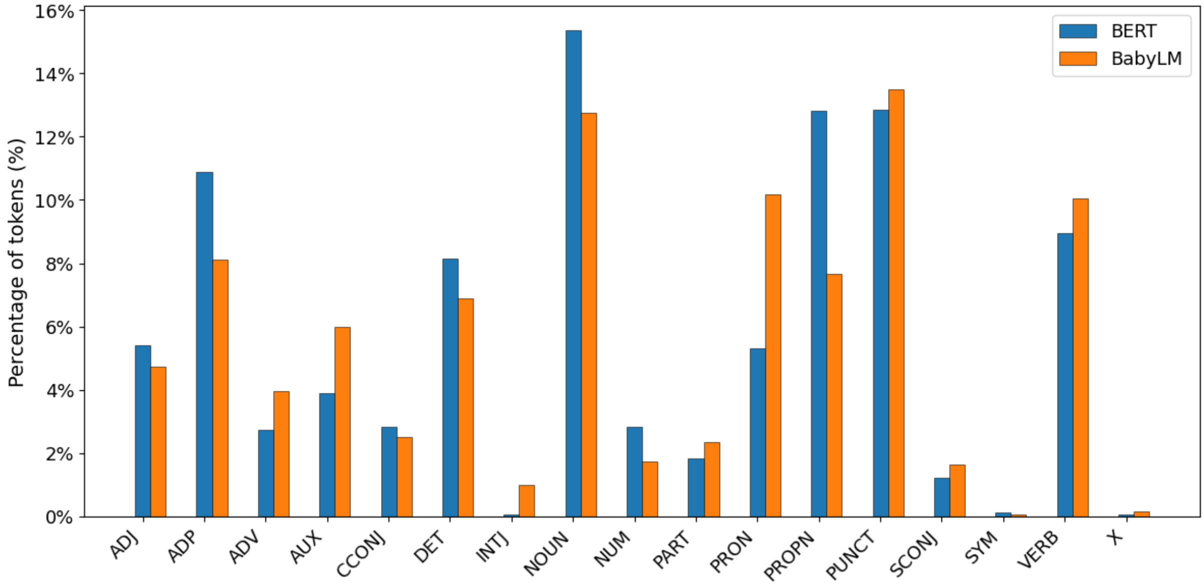

Figure 11: Part-of-Speech distribution across corpora

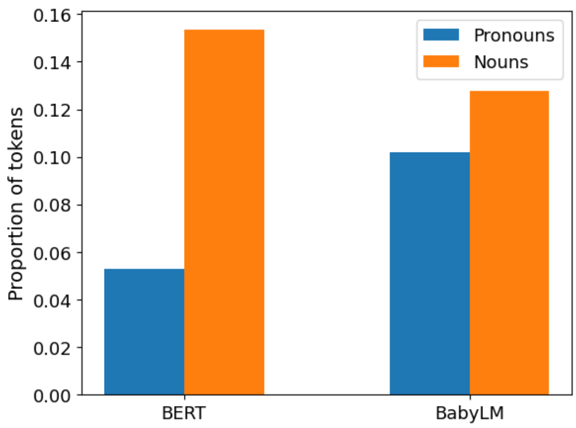

Figure 12: Pronoun vs. noun distribution across corpora

pus between 5th and 6th grades, which corresponds to students aged 10-12 and thus aligns with what would be expected for BabyLM.

Unsurprisingly, when investigating the type-token ratio shown in Table 9 to ascertain the richness of the vocabulary, we see that BERT has the more diverse vocabulary. BabyLM is again much simpler. This is likely due to its enforced child-alignment.

D.1.3 Coherence

Finally, we use the first-order coherence to examine how tightly each text corpus adheres to a topic. In Table 10, we see that the corpora score highly, with both being similarly consistent.
Overall, the BabyLM corpus is markedly simpler in text complexity yet maintains coherence comparable to BERT. Consequently, if it contains sufficient bias-inducing material, it is likely adequate to impart those biases to the model.

D.2 Toxicity, Sentiment, and Emotions

Furthermore, we need to understand the tone of the corpora and the context in which they present their different topics. While no exact link has been created, there is some evidence that displaying information in a negative or toxic light pushes the model to develop stronger biases (BigScience-Workshop et al., 2022). As such, we must examine the presence of toxicity, hate-speech, sentiment, and the overall emotional composition throughout the corpora.

D.2.1 Emotion Scores

Starting with emotion detection, we are interested in measuring Ekman's core emotions: joy, sadness, fear, surprise, anger, and disgust (Ekman et al., 1999). To this end, we take advantage of the *NRC Emotion Lexicon*, which links individual words to various emotions (Mohammad and Turney, 2010), allowing us to infer the overall emotional tone of the corpus. The normalised *emotion score* of a corpus is calculated by dividing the number of words expressing a specific emotion by the total number of eligible words. This *emotion score* is computed for all six emotions across all our corpora.

The results of the lexical analysis can be seen in Figure 13. While the emotions are mostly balanced across the corpora, BabyLM is consistently

Corpus	Avg. syllables per word	Avg. word length	Avg. sentence length	FKGL
BabyLM	1.25	4.01	15.91	5.40
BERT	1.38	4.57	19.84	8.43

Table 8: Readability and length statistics for the BERT and BabyLM corpora.

Corpus	Type–token ratio
BabyLM	0.33
BERT	0.53

Table 9: Type–token ratio for the BERT and BabyLM corpora

Corpus	First-order coherence
BabyLM	0.811
BERT	0.802

Table 10: First-order coherence scores for the BERT and BabyLM corpora

more emotional, especially in terms of joy and surprise. However, in both cases, the difference is mild and both are positive emotions, which are not the carriers of bias.

D.2.2 Sentiment

Works like the one by Köksal et al. (2023) have indicated that sentiment is passed from corpora into the biases of models, forging positive or negative connections that might end up strengthening stereotypes. Because of this, we measure the sentiments in each corpus, both the overall ones and those connected to specific topics.

We use the established RoBERTa-based solution by Barbieri et al. (2020), which is one of the most popular publicly available sentiment analysis models. With it, we classify each corpus as positive, negative, or neutral. Finally, we compute the *sentiment score* in the same fashion as we computed the *emotion score*, only switching from word-based to sentence-based evaluation.

Figure 14 shows the overall sentiment profile of the corpora. The BabyLM corpus is consistently less neutral, reinforcing that it is well-positioned to transfer the biases into the model despite its child-alignment.

Examining the results more closely, Figure 15 displays the sentiment of sentences containing a bias-inducing word from a specific category for each corpus. To calculate this, we take all eligible sentences per topic and subtract the percentage of negative sentences from the percentage of positive sentences. Here, we can see that the BabyLM corpus remains overall more negative than the BERT corpus, with the exception of the Muslim topic. While it carries more negativity, the ratios between topics are similar, meaning that it retains the same bias orientation. The only true divergence is higher negativity towards LGBTQ+ topics.

In summary, BabyLM proves to be more emotional and negative than BERT, while retaining the same larger trends, meaning that the corpora are largely compatible. The evidence supports the BabyLM corpus's ability to transmit biases to a model trained on it.

D.3 Specifications Regarding Toxicity

While this topic has already been covered in Section 4.1, we want to provide a validation of the result. Given that we have established that the BabyLM corpus mostly contains simple sentences, the toxicity and hate-speech results in Table 4 raise the question of whether the toxicity and hate-speech labels are overly sensitive. To investigate this, we sampled 200 such sentences, identifying that the vast majority are indeed toxic, containing slurs or highly aggressive expressions. For illustration, we list some sampled examples in Table 11 (**Contains explicit offensive statements**). These examples display blatantly racist and sexualised terms, showing that, despite its child-alignment, BabyLM contains diverse and toxic texts.

Sentence
Her earhole ain't big enough for f***ing!
cause they don't want to look like morons too.
He was a black n****r what?

Table 11: Sampled examples of toxic sentences from the BabyLM corpus

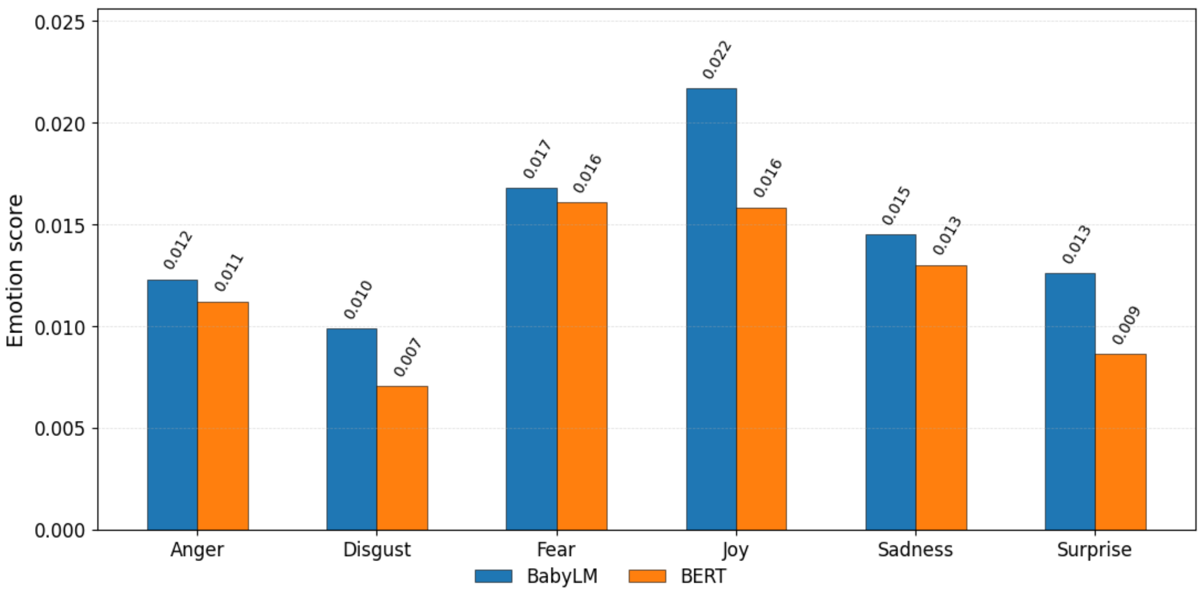

Figure 13: Distribution of emotions across the corpora using the emotion score obtained by lexical analysis

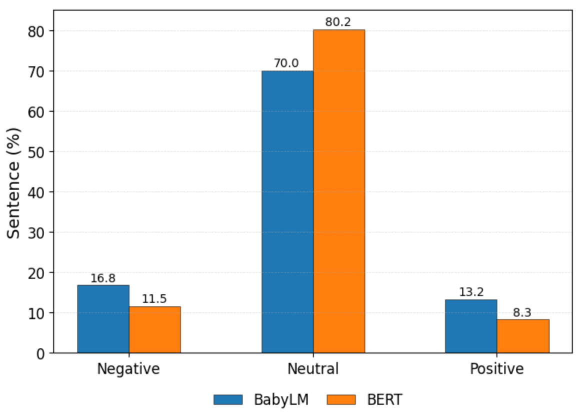

Figure 14: Distribution of sentences per sentiment across corpora

E INLP Reduces Performance Penalty for Over-Fitted BabyLMs

The INLP experiment has also revealed an important piece of behaviour from LTG-BERT that concerns the BabyLM challenge itself. As mentioned, INLP helped improve LTG-BERT's performance, likely by reducing the issues connected to overfitting on its training corpus.

This is notable because LTG-BERT is the best-performing BabyLM-class MLM, with the SOTA BabyLM being GPT-BERT, which was trained with the same training data-to-epoch ratio. Thus, if INLP debiasing makes LTG-BERT surpass the SOTA model on our performance tasks, as shown in Figure 16, it could even yield further improvements to other over-fitted BabyLMs, generating further performance improvements.

F Implementing LTG-Baseline Pre-training

We must define the exact specifics of training our own LTG-Baseline model. This requires us to prepare the corpus, define a training script, and specify the exact hyperparameters. Following the specifications from the authors (Hu et al., 2024b), we use the predefined BabyLM corpus established in Section 4.1, and adopt the LTG architecture by Samuel et al. (2023) as the starting point. However, an issue with further implementation is that the authors have not published the LTG-Baseline training script or its hyperparameters.

To resolve this, we reached out directly to the authors, who supplied us with all the necessary details. They used the standard MLM training script by Wolf et al. (2020) with hyperparameters, which are provided in Table 12.

Furthermore, for purposes of model comparison, epochs cease to be a representative unit when we alter the corpus. To remedy this, we note that, in the standard training, 20 epochs translate to roughly 140,000 training steps. Thus, we compare all of our models on the interval from 0 to 140,000 steps, removing any unfair advantage stemming from an expanded corpus.

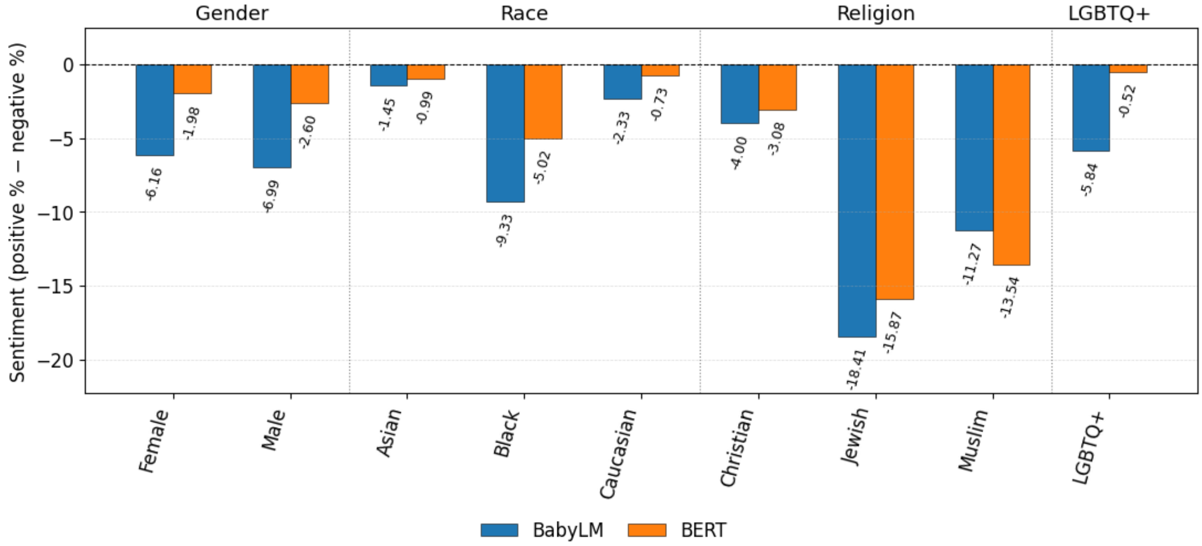

Figure 15: Sentiment stances to the selected topics across corpora

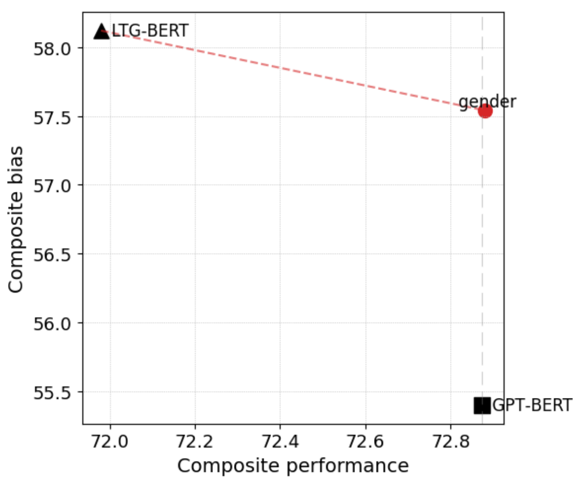

Figure 16: Performance improvement of INLP-debiased LTG-BERT compared with the GPT-BERT BabyLM

Hyperparameter	Value
max_seq_length	128
per_device_train_batch_size	128
num_train_epochs	20
learning_rate	5e-4
adam_beta1	0.9
adam_beta2	0.999
adam_epsilon	1e-8
max_grad_norm	1
warmup_steps	0

Table 12: LTG-Baseline pre-training hyperparameters

G LLM-in-the-loop Debiasing Implementation

The LLM-in-the-loop detoxification approach has garnered interest because it allows us to remove toxicity without removing information from a corpus (Yuan et al., 2025). The method uses an LLM to rewrite any toxic sentence in a non-toxic way whilst preserving the sentence's meaning. Thus, we are able to remove toxicity without disrupting coherence.

For the purposes of implementation, we selected Llama-3.3-70B to detoxify the sentences (Grattafiori et al., 2024). It was picked due to being a widely used and high-performing open-source model. In order to identify a well-performing detoxification prompt, we sampled 20 toxic sentences and tested multiple prompt variants, with the selected one, together with an example, shown in Figure 17.

H Perturbation Augmentation Implementation

The Perturbation Augmentation debiasing technique by Qian et al. (2022) relies on the idea of using a pre-trained LLM (perturber) to rewrite the corpus, randomly swapping every demographic reference for a different one.

In practice, this means that when supplied with a chunk of text, a target word, and a target topic, the perturber changes the gender, age, or race of the subject in the chunk to a new one randomly selected from the set list in Figure 13. If applied to

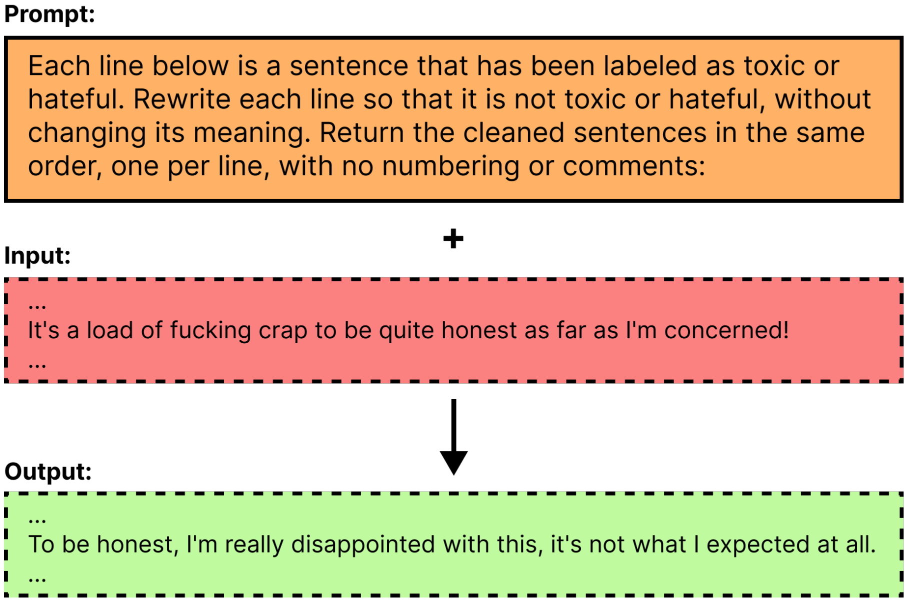

Figure 17: Detoxification prompt and its effect on an example

the entire corpus, this ensures uniform distribution of topics. An example of this kind of perturbation is shown in Figure 18.

To apply this perturbation to the BabyLM corpus, we use the perturber trained by the authors, ensuring the quality of the result. Unfortunately, we also need the target words, which the authors have not provided. Nevertheless, we resolved this by extracting a noisier list of target words from the perturber's training data.

During the perturbation process, we split the corpus into 128-token chunks, extract all target words from the chunk, randomly select one of them together with a random sub-category, and perturb the chunk. We only focus on race and gender perturbations as age proved to generate more erroneous changes. If the chunk does not change, we repeat this process with another word until we change the chunk or exhaust the target words. Thus, the noisy ineffective words do not disrupt the experiment. With the entire corpus transformed, Table 13 details the distribution of changes.

Category	Subcategory	Perturbed Chunks
Overall Corpus	Any Change	84.9%
	Gender	79.0%
	Race	5.9%
Gender Perturbation	Non-binary	27.9%
	Woman	26.8%
	Man	24.2%
Race Perturbation	Pacific-Islander	1.1%
	Native-American	1.0%
	White	1.0%
	Asian	1.0%
	Black	1.0%
	Hispanic	1.0%

Table 13: Change distribution in the perturbed corpus

I Generative AI

An AI assistant was used to help with grammar and language editing.

J Computational Resources

For all debiasing experiments, covering both fine-tuning and pre-training, we utilised a server with four A100 GPUs. Most evaluation tasks were conducted on a single NVIDIA T4 GPU. In total, including test runs, the experiments consumed approximately 700 GPU-hours. No hyper-parameter search experiments were conducted for any of the experiments.

K Code Repository

The code is available in a dedicated GitHub repository.[1]

[1] https://github.com/trhlikfilip/bias-dynamics-sandbox

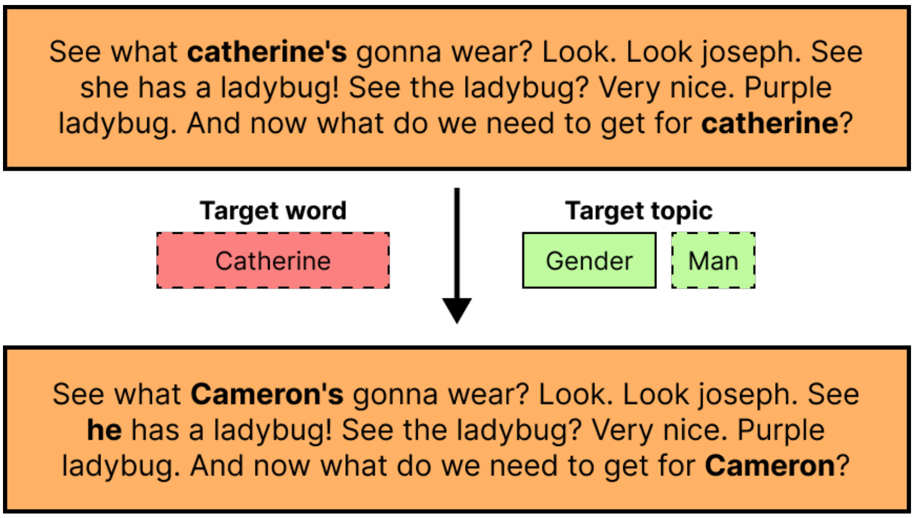

Figure 18: Perturbation showcase