Are you talking to ['xem'] or ['x', 'em']? On Tokenization and Addressing Misgendering in LLMs with Pronoun Tokenization Parity

Anaelia Ovalle[*‡] **Ninareh Mehrabi**[†] **Palash Goyal**[†]

Jwala Dhamala[†‡] **Kai-Wei Chang**[‡†] **Richard Zemel**[†]

Aram Galstyan[†] **Yuval Pinter**[†] **Rahul Gupta**[†]

[‡]University of California, Los Angeles [†]Amazon Alexa

Abstract

A large body of NLP research has documented the ways gender biases manifest and amplify within large language models (LLMs), though this research has predominantly operated within a gender binary-centric context. A growing body of work has identified the harmful limitations of this gender-exclusive framing; many LLMs cannot correctly and consistently refer to persons outside the gender binary, especially if they use neopronouns. While data scarcity has been identified as a possible culprit, the precise mechanisms through which it influences LLM misgendering remain underexplored. Our work addresses this gap by studying data scarcity's role in subword tokenization and, consequently, the formation of LLM word representations. We uncover how the Byte-Pair Encoding (BPE) tokenizer, a backbone for many popular LLMs, contributes to neopronoun misgendering through out-of-vocabulary behavior. We introduce *pronoun tokenization parity* (PTP), a novel approach to reduce LLM neopronoun misgendering by preserving a token's functional structure. We evaluate PTP's efficacy using pronoun consistency-based metrics and a novel syntax-based metric. Through several controlled experiments, finetuning LLMs with PTP improves neopronoun consistency from 14.5% to 58.4%, highlighting the significant role tokenization plays in LLM pronoun consistency.

1 Introduction

Gender bias in natural language processing (NLP) has been widely studied in the context of binary gender, however mitigating harmful biases for underrepresented gender minorities remains an active area of research Sun et al. [2019], Stanczak and Augenstein [2021]. As demonstrated in prior research Dev et al. [2021], Ovalle et al. [2023], Hossain et al. [2023], large language models (LLMs) struggle with correctly addressing individuals using non-binary pronouns, especially when they are neopronouns (e.g., *xe*, *ey*)[2]. In addition to their respective findings, these works necessarily highlight the connection between LLM misgendering and data scarcity; neopronouns are often severely underrepresented in a pretraining text corpus, thus impacting the LLMs ability to learn how

[*]Corresponding Author, anaelia@cs.ucla.edu

[2]https://nonbinary.wiki/wiki/English_neutral_pronouns

37th Neural Information Processing Systems Queer in AI Workshop (NeurIPS Queer in AI 2023).

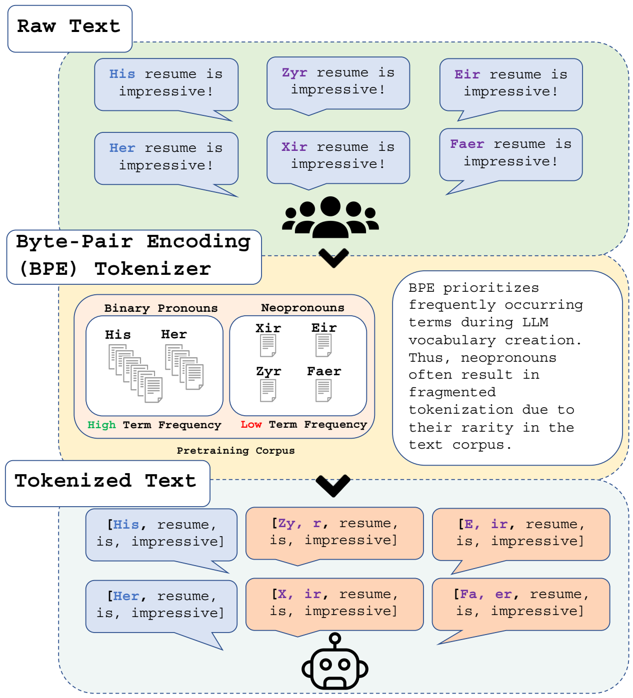

Figure 1: LLMs critically rely on their tokenizer's predefined base vocabulary to form their representations. As BPE prioritizes constructing full tokens for frequently occurring terms, neopronoun sparsity in a text corpus impacts tokenization and its subsequent LLM representation. Neopronouns are fragmented into subword units while binary pronouns are represented as single tokens, thereby introducing syntactic challenges for LLMs which we investigate in this paper.

to use them appropriately. However, investigation into the specific mechanisms by which data scarcity impacts LLM misgendering behavior remains limited. Our study tackles this research gap from a new angle, focusing on a fundamental aspect of LLM representation development: tokenization. Figure 1 illustrates how binary pronouns and the neopronoun *xir* are tokenized under Byte-Pair Encoding (BPE), today's most common subword tokenizer used in popular LLMs such as GPT-2 Radford et al. [2019], GPT-3 Brown et al. [2020], and LLaMA Touvron et al. [2023]. While binary pronouns (*her* and *his*) are tokenized as single units, neopronouns like *xir* and *eir* are fragmented into two subword tokens due to their infrequency within the tokenizer's training corpus. This out-of-vocabulary (OOV) tokenization subsequently forces the LLM to rely on granular subword tokens to learn the neopronoun's representation (embedding). As these subwords are present in multiple words, their embeddings incorporate information from these common words, making it challenging to distinguish the neopronoun in the model. This paper explores the impact this low-resource tokenization has on an LLM's ability to correctly refer to a person's neopronouns (i.e. pronoun consistency). Our experiments reveal that this fragmentation significantly impairs an LLM's ability to correctly and consistently use neopronouns.

Guided by prior NLP literature detailing LLM syntactic challenges introduced by OOV, our paper introduces a novel mitigation strategy termed *pronoun tokenization parity* (PTP). PTP centers aligning, or establishing parity, between neopronoun and binary pronoun tokenization. We preserve the neopronoun's functional structure as a pronoun by representing it as a single token for LLM input. We evaluate the efficacy of PTP both with typical pronoun-consistency metrics alongside a novel syntactic knowledge-based metric strongly associated with pronoun consistency. Furthermore, given the substantial training costs of LLMs and their resulting environmental impact, we present a cost-effective alternative which exploits an LLM's existing grammatical knowledge to achieve substantial improvements in LLM pronoun consistency.

Finetuning GPT-based models across carefully augmented neopronoun datasets show PTP providing up to 58.4% pronoun consistency, compared to 14.5% when traditionally finetuning without PTP (§5.1). When finetuning only the LLM's lexical layer, a technique commonly seen in multilingual NLP for cross lingual transfer, it surpasses the performance of full fine-tuning for most of our models

Pronoun Case	He	She	Xe
Nominative	[he]	[she]	[xe]
Accusative	[him]	[her]	[x, em]
Pronominal Possessive	[his]	[her]	[x, ir]
Predicative Possessive	[his]	[hers]	[x, irs]
Reflexive	[him, self]	[her, self]	[x, ir, self]

Figure 2: BPE Tokenization of Binary Pronoun Cases and Neopronoun Cases for *Xe*.

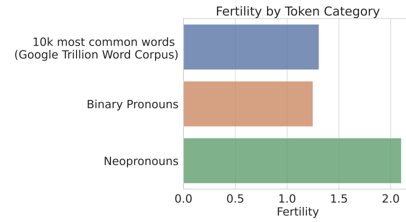

Figure 3: Fertility by pronoun category. Details and definitions of measurement found in §A.2.

(75%) while simultaneously reducing training time by up to 21.5% (§5.2). Testing our methods at scale, we find that lexical fine-tuning consistently improves LLM pronoun consistency across model sizes, with smaller models experiencing the most improvements - even matching the performance of models twice their size (§5.3).

2 Background

Gender-Inclusive NLP Gender biases have been studied across several NLP contexts, including machine translation Stanovsky et al. [2019], coreference resolution Rudinger et al. [2018], Zhao et al. [2018], and named entity recognition Mehrabi et al. [2019]. Recent works expand gender bias evaluations to harms unique to non-normative gender communities within LLMs Dev et al. [2021], Hossain et al. [2023], Ovalle et al. [2023], Nozza et al. [2022], Felkner et al. [2023], of QueerInAI et al. [2023]. Dev et al. [2021] examines non-binary gender biases in static and contextual language representations, highlighting how data limitations affect these embeddings. Similarly, Ovalle et al. [2023] explores misgendering and harmful responses related to gender disclosure using their `TANGO` framework, pointing to challenges in neopronoun consistency, possibly due to data scarcity. Hossain et al. [2023] corroborates these findings with an incontext-learning evaluation and analyses into LLM pretraining corpus statstics. Despite exploring various in-context learning strategies, they find persistent gaps between binary pronoun and neopronoun misgendering.

While these studies collectively emphasize data scarcity's impact on neopronouns, questions remain regarding the precise mechanisms through which data scarcity shapes neopronoun representations and subsequent LLM pronoun consistency. In this study, we investigate the pivotal role of BPE tokenization due to its inextricable link to both data scarcity and the construction of LLM word representations. We see how studying BPE behavior in resource-constrained settings yields valuable clues into LLM misgendering, offering a new approach towards addressing this issue.

BPE Tokenization Byte-Pair Encoding (BPE) is a subword tokenization technique Sennrich et al. [2016] that constructs token vocabularies by iteratively merging frequently occurring adjacent token pairs until a predefined vocabulary size is reached. BPE relies on token frequency in the corpus to create these merge rules and does not assign a specific "unknown" token `[UNK]` to unseen words. In cases where a word is not present in the vocabulary, BPE decomposes it into subword units, with the most granular unit being individual characters. This method is designed to encompass both frequent and infrequent subword units, yet it remains solely influenced by text frequency and does not take into account inductive or semantic factors.

3 Low-Resource Challenges for BPE & LLMs

The Out-of-Vocabulary Problem The Out-of-Vocabulary (OOV) problem refers to when an LLM encounters a word outside its predefined vocabulary, due to novelty or rarity in the training corpus. For LLMs using subword tokenizers like BPE, OOV words are broken down into multiple smaller tokens. BPE tokenization seems to reflect this OOV behavior with neopronouns like *xem*. Shown in Figure 2, the BPE tokenizer breaks *xem* down into character and subword token. Unlike tokenization for binary pronouns *he* and *she*, BPE does not treat it as single unit, indicating that the token cannot be constructed by the LLM's predefined vocabulary Yehezkel and Pinter [2022]. This phenomenon presents numerous challenges across NLP tasks, with notable documentation in machine translation Domingo et al. [2018], Huck et al. [2019], Araabi et al. [2022].

OOV and Syntactic Knowledge Gaps in LLM syntactic knowledge due to OOV behavior is well documented. Wang et al. [2019] highlight OOVs' detrimental impact on part-of-speech (POS) discernment, resulting in high error rates for OOV words. Such challenges manifest across tasks reliant on syntactic understanding, such as POS tagging Wicaksono and Purwarianti [2010], Pinter et al. [2017], Wang et al. [2019], named entity recognition Dařena and Suss [2020], Wang et al. [2022], and quality estimation for machine translation Domingo et al. [2018], Huck et al. [2019], Araabi et al. [2022]. The authors stress token-level comprehension's importance in these tasks, underscoring the need to address OOV challenges for improved downstream performance.

Rust et al. [2021] also highlights LLM sensitivity to tokenization through a comprehensive empirical analysis of token fragmentation across several languages. They introduce 'fertility', the average number of subwords produced per tokenized word, to quantify tokenizer impacts on LLM performance across benchmark NLP tasks: the closer fertility is to 1, the better the tokenizer performs. After undergoing BPE tokenization, neopronouns frequently decompose into multiple subword units, resulting in elevated fertility scores. These fertility scores are substantially higher than those found for binary pronouns, as depicted in Figure 3. Given the observed LLM sensitivity to tokenization reflected in POS tagging and dependency parsing by Rust et al. [2021], we posit that these decompositions adversely impact an LLM's ability to handle neopronouns.

OOV and Context Formation LLMs rely on context to make predictions and generate coherent text. Linguistic ambiguity may arise when splitting *xem* into the tokens *x* and *em*, reflected in challenges to use these subword units in their appropriate context. For instance, if pretraining data shows subtokens *x* and *em* as (parts of) nouns or abbreviations, using these tokens as pronouns reflects a less familiar context for the LLM Dev et al. [2021], thereby introducing confounding signals from many different contexts into one token. Furthermore, *xem* combines the relatively rare character *x* and a frequently occurring character pair *em*. Since frequency in text plays a significant role in neopronoun representation both for the tokenizer and the LLM's context formation, contextual complexity may be introduced when LLMs see these tokens used in ways that do not reflect their dominant contexts Pinter et al. [2017].

4 Pronoun Tokenization Parity

English pronouns serve as building blocks for language acquisition. Termed *functional morphemes*, these small, self-contained units of meaning reflect specific English grammatical functions Fortescue [2005], Penny Eckert and Ivan A. Sag [2011]. Works have shown that BPE captures varied morphological segments in text Park et al. [2020], Bostrom and Durrett [2020], with BPE subword organization being a reflection of morphological complexity Gutierrez-Vasques et al. [2021]. This unit of meaning seems to be captured for binary pronouns in BPE tokenization, reflected in the fact that it is tokenized as one whole, atomic token. However, unlike binary pronouns, neopronouns like *xem* are not tokenized as functional whole-word tokens but rather divided into subword units. We posit that this absence of functional preservation hinders the syntactic coherence of neopronouns in LLMs, thereby influencing pronoun consistency. As such, this section offers a mitigation approach informed by these fundamental insights.

4.1 Neopronoun BPE Tokenization

In order to improve LLM neopronoun consistency, we introduce *pronoun tokenization parity*, or PTP, to preserve a token's functional integrity during BPE tokenization. By aligning this approach with binary pronoun tokenization, we hypothesize that this will improve an LLM's grammatical representation of neopronouns, thereby improving a model's ability to use them properly. We produce neopronouns as cohesive linguistic units through the use of **special tokens**. In LLMs, special tokens may offer functionality such as padding (`[PAD]`) or sentence boundary marking (`[EOS]`), but custom definition of others is possible. We allocate a special token to each neopronoun, allowing it to be represented as a single token, similarly to binary pronouns. We posit that this equivalence will improve the LLM's capability to capture neopronoun functional morphemes, thereby improving its ability to recognize them as pronouns. Additional details and instructions for reproducing PTP are located in Appendix 1.

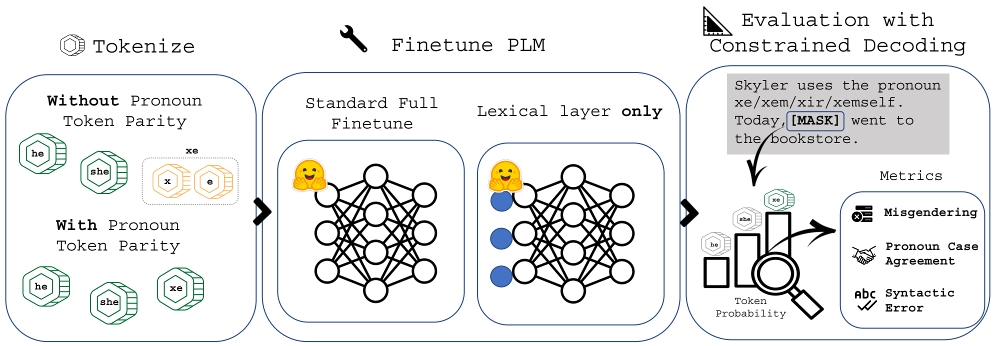

Figure 4: We (1) tokenize neopronouns using PTP for a given LLM, (2) either fully finetune or only finetune the LLM lexical layer with data containing neopronouns, and (3) determine our method's efficacy in reducing LLM misgendering using constrained decoding approach across 3 metrics.

To operationalize PTP, we employ two finetuning paradigms with an open-source LLMs varying in capacity and neopronoun data scarcity. We provide an overview of our framework in Figure 4. The model and data details, along with our performed experiments, are detailed in §5. Formally, we extend the pretrained token embeddings of a transformer-based LLM $E_1^{\text{orig}}, E_2^{\text{orig}}, \ldots, E_n^{\text{orig}}$, where n represents the vocabulary size of the original model. We introduce new embeddings E^{PTP} for each of m unique pronouns in the set of neopronoun cases (i.e., pronoun family) S, resulting in an extended vocabulary: $\{E_1^{\text{orig}}, \ldots, E_n^{\text{orig}}\} \cup \{E_1^{\text{PTP}}, \ldots, E_m^{\text{PTP}}\}$.

4.2 Metrics

English pronouns must agree with their subject in gender, case, and number Garner [2016]. To quantify a model's ability to understand various pronoun forms, we outline three metrics, two of which are derived from these grammar rules and the third is novel metric which we introduce in this paper: *pronoun consistency* (PC) to assess pronoun-gender agreement; *case agreement* (CA) to test agreement with pronoun case; and *syntactic perturbation error* (SE) to assess syntactic robustness to word insertion adversarial attacks which render a sentence grammatically incorrect or change the sentence's meaning. PC is the most important metric to determine improvement in pronoun-gender agreement, whereas CE & SE provide further insight into sources of error. These metrics are employed in a constrained decoding setting, consistent with Hossain et al. [2023]. Given a masked template, the LLM predicts the most likely pronoun given a pool of pronouns of the same form. We provide further dataset details in Appendix A.5.

Pronoun Consistency Let S be a set of unique pronoun families with $|S|$ pronoun families. Each pronoun family $M \in S$ contains $|M|$ English pronoun forms. Within a collection of masked templates T, `[MASK]` is replaced with a pronoun $p \in M$ for all $M \in S$, resulting in the filled template set T^*. In line with Hossain et al. [2023], each template starts with a person's name and their pronoun declaration (i.e. nominative/accusative/genitive/ reflexive), followed by a sentence containing a `[MASK]` token which expects a pronoun. For example: *Casey uses the pronouns he/him/his/himself. Upon recognizing Casey, the fan asked [MASK] for an autograph.*. For a template t consisting of m tokens $x_1, x_2, \ldots, x_m$, the token generated at `[MASK]`, is determined by `argmax` transition probability from the pronoun pool.

$$\hat{y}_t = \text{argmax}_{p \in S} P(x_i = s | x_{<i}) \tag{1}$$

.

We denote the set of filled templates as C. Each filled template is then compared to its golden label example $c \in C^*$, containing the correct pronoun for that template-name-declaration combination.

To evaluate pronoun consistency, we compare the model's chosen pronoun for a template, $\hat{y}_t$, to the template's correct pronoun, y_c, and then calculate the accuracy over all templates:

$$\frac{1}{|T^*|} \sum_{t \in T^*, y \in C^*} \delta(\hat{y}_t, y_c). \tag{2}$$

Case Agreement Evaluating case agreement is essential to assess a model's pronoun usage proficiency. Ideally, an LLM would generate case-agreeing sentences like "She went to the store." instead of "Hers went to the store." To evaluate this, we use the same approach as above, instead focusing on assessing expected versus predicted pronoun case for a given pronoun family.

We cannot rely on transition probabilities to determine grammatical correctness, as they are conditioned only on previous tokens. For example, a sentence like "Casey went to the store for `[MASK]` mom" can have its mask replaced with "her" or "herself" and still be grammatically correct, as it only considers the previous tokens during inference. Therefore, we acquire the model's predicted output across all pronoun cases for a given family $s \in Q$, minimizing its loss (i.e., maximized probability). We then calculate the same accuracy using Eq. (2).

$$\text{argmin}_{s\in Q}\left(-\sum_{i=1}^{N}\log P_\theta(x_i|x_{<i})\right). \tag{3}$$

Syntactic Perturbation Error Recent work finds that LLM-generated text prompted with neopronouns frequently produces ungrammatical text. Ovalle et al. [2023] finds that such text prefixes neopronouns with articles and determiners (e.g., 'the', 'a', 'these'). Further quantifying to what extent misgendering correlates with poor grammar highlights potential avenues for amelioration. For instance, Bao and Qiao [2019] find that improving an LLM's syntax can help binary pronoun resolution in low-resource settings.

To assess an LLM's grammaticality with respect to neopronouns, we introduce a measurement based on adversarial word insertion attacks that mimic the observed ungrammatical behavior found in Ovalle et al. [2023]. Similar to pronoun consistency, we use LLM transition probabilities to evaluate whether and to what extent they tend to use neopronouns in ungrammatical ways, providing insight into the model's syntactic understanding of them. We use the same templates as before, but now we augment each `[MASK]` as `[DET][MASK]`, where `[DET]` is replaced by singular and plural determiners (e.g., 'this', 'those', 'these'), articles (like 'the', 'a'), or no determiner at all. Example templates are found in subsection A.5. Next, we analyze the LLM's output by calculating the `argmax` of the transition probability for all potential substitutions of `[DET]`. Ideally, an LLM that correctly uses neopronouns will choose a template that does not include a determiner.

5 Experiments

In the following sections, we conduct controlled experiments to assess the impact of PTP on LLM neopronoun consistency. We compare PTP performance to original BPE tokenization across different resource settings, shedding light on the role tokenizers play in LLM pronoun consistency gaps. Additionally, we explore resource-efficient mitigation techniques and evaluate the scalability of our methodology, providing insights into scenarios with the most substantial improvements.

We focus on using PTP for the neopronoun family *xe*, for several reasons. First, *xe* ranks among the most widely adopted non-binary pronouns Gender Census [2023]. Second, it is well-documented that non-binary pronouns exhibit a diverse range of linguistic variations, spanning from closed to open word class forms Miltersen [2016], Lauscher et al. [2022]. This diversity requires a nuanced yet flexible approach. By focusing on the *xe* pronoun family, we showcase the effectiveness of PTP while providing a generalizable framework for researchers to build off for studying non-binary pronouns within their respective linguistic contexts.

Models We conduct our experiments using the Pythia model suite.[3] We choose this framework as it parallels state-of-the-art architecture; Pythia models all consist of a GPT-Neo-X architecture, an open-source alternative to GPT-3 models. Notably, it is based on a BPE tokenizer Biderman et al. [2023] and trained on a commonly accessible dataset, the PILE. Furthermore, as these models vary from 70M to 12B parameters, they provide an ideal environment for investigating LLM knowledge development with PTP across model capacity.

Datasets Due to the limited availability of textual data containing neopronouns, we make use of the Wikibios dataset,[4] which consists of narratives about real individuals. To mitigate this scarcity,

[3] `https://github.com/EleutherAI/pythia`

[4] `https://huggingface.co/datasets/wiki_bio`

Model	Pronoun Consistency (↑) He	She	Xe	Case Agreement (↑) He	She	Xe	Error (↓) He	She	Xe
$T_{\text{Orig}} + M_{\text{Baseline}}$	0.968	0.716	0.007	0.677	0.607	0.200	0.238	0.169	0.850
$T_{\text{Orig}} + M_{\text{Full}}$	0.896	**0.861**	0.145	0.685	0.597	0.200	**0.239**	**0.168**	0.895
$T_{\text{Orig}} + M_{\text{Lex}}$	0.865	0.729	0.168	0.617	0.572	0.200	0.290	0.232	0.654
$T_{\text{PTP}} + M_{\text{Full}}$	**0.948**	0.835	0.378	**0.695**	**0.611**	0.294	0.278	0.210	**0.270**
$T_{\text{PTP}} + M_{\text{Lex}}$	0.850	0.722	**0.536**	0.632	0.591	**0.325**	0.258	0.218	0.348

Table 1: 70M-parameter model results at 10% data resource level. T_{Orig}= without `PTP`, T_{PTP}= with `PTP`, M_{Full}= full finetuning, M_{Lex}= only lexical finetuning. M_{Baseline}= original model, no finetuning.

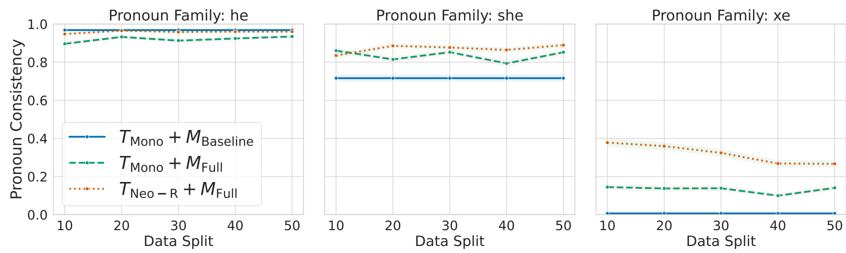

Figure 5: 70M-parameter model with full finetuning across data resource sizes. Plot shows using PTP most effective in maintaining pronoun consistency for pronoun family *xe*.

we employ a counterfactual data augmentation technique that replaces a variable fraction of binary pronouns with their neopronoun equivalents. We do this for two primary reasons. Firstly, individuals who use neopronouns often have historical associations with binary pronouns. Recognizing this and considering the substantial shortage of neopronoun representation in pretraining corpora, we aim to incorporate narratives that resonate with a broader spectrum of individuals Talat and Lauscher [2022]. This approach enables LLMs to learn neopronoun usage within more comprehensive, diverse, and real-world contexts. Secondly, the dataset is readily available via HuggingFace, allowing for ease in reproducibility. We filtered the Wikibios dataset, comprised of 728,321 English biographical texts from Wikipedia, to retain texts containing binary pronouns, resulting in 462,345 examples. We replaced each binary pronoun with its corresponding case for *xe*, incorporating correct possessive forms using the spaCy part-of-speech tagger [Honnibal et al., 2020].

To evaluate PTP, we conducted experiments fine-tuning models on this augmented data with varying amounts of neopronouns, representing different resource levels: 10%, 20%, 30%, 40%, and 50%. In the 50% setting, the dataset is evenly split between *xe* and binary pronouns. No biography text appears more than once in the dataset splits. For evaluation, we utilized the `MISGENDERED` dataset by Hossain et al. [2023], containing added templates and names from `TANGO` [Ovalle et al., 2023] (more details in §A.5), resulting in 93,600 templates for evaluation across our defined metrics.

5.1 Experiment 1: How helpful is Pronoun Token Parity for reducing misgendering in LLMs?

We assess PTP's impact on reducing misgendering for a compact 70M parameter Pythia model. We prepare two versions for fine-tuning: one without PTP (original tokenizer, T_{Orig}) and one with PTP (T_{PTP}). PTP embeddings for T_{PTP} are initialized with a random Gaussian (μ=0 and σ=0.02). M_{Full} encompasses all models fine-tuned with standard full finetuning, and M_{Baseline} represents the HuggingFace baseline checkpoint, utilizing the non-PTP tokenizer T_{Orig}. Fine-tuning is done for five epochs with a batch size of 128, a learning rate of 10^{-4}, and early stopping based on cross-entropy loss in the validation set with a patience of 2. To expedite model training and inference, all models undergo fine-tuning using FP16 mixed precision and two gradient accumulation steps. We provide further details on our setup in §A.4.

Results As results in Table 1 show, both $T_{\text{PTP}} + M_{\text{Full}}$ and $T_{\text{Orig}} + M_{\text{Full}}$ demonstrate an improved neopronoun consistency over the baseline 70M Pythia model. This improvement is expected, considering their increased exposure to neopronouns during fine-tuning. Interestingly, models using a tokenizer with PTP consistently outperformed those without PTP, as shown in Figure 5. Improvement over this baseline is observed across data resource levels, especially at lower resource levels, where $T_{\text{PTP}} + M_{\text{Full}}$ more than doubles the consistency improvements from $T_{\text{Orig}} + M_{\text{Full}}$ (37.8% vs. 14.5%). Binary pronoun consistency remains stable with PTP, with $T_{\text{PTP}} + M_{\text{Full}}$ even enhancing *she* pronoun consistency. Our findings indicate that a significant portion of neopronoun consistency disparities can be attributed to OOV tokenization due to neopronoun scarcity, motivating further investigation into potential enhancements to tokenization.

5.2 Experiment 2: Can a more resource-efficient approach to PTP still reduce LLM misgendering?

Full finetuning can be resource-intensive, posing challenges for machine learning researchers due to the associated time and compute costs. Since Pythia has already learned English syntax and binary pronouns from its pretraining, we hypothesize that full finetuning may not be necessary to learn new neopronouns following English grammar rules. Inspired by cross-lingual transfer techniques Artetxe et al. [2019b], de Vries and Nissim [2020], we experiment with finetuning only Pythia's lexical embedding layer, leaving transformer weights unchanged. Unlike Artetxe et al. [2019b], we avoid training the transformer weights after freezing lexical embeddings since the new tokens already conform to English grammar and syntax, eliminating the need for the transformer to adapt to a different language. Additionally, differing from the approach by de Vries and Nissim [2020], we avoid resetting the entire lexical embedding layer to retain the prelearned English grammar dependencies. We compare models that only had their lexical layers finetuned, M_{Lex}, with a baseline where a non-PTP model's lexical layer is finetuned. We follow the same setup as before but increase the learning rate to 10^{-3} to encourage more rapid adaptation to the new vocabulary.

Results As shown in Table 1, employing PTP with lexical finetuning outperformed standard full finetuning with the original BPE tokenizer (53.6% vs. 37.8%). T_{PTP} + M_{Lex} also outperformed T_{Orig} + M_{Lex} (53.6% vs 16.8%), further supporting that benefits come from the PTP, rather than the lexical finetuning alone. Notably, M_{Lex} consistently outperformed finetuning across various data resource levels (see Appendix Table 7a). Given identical context-dependent structures between pronouns, these results demonstrate pronoun adaptation via *only* learning a new context-independent lexical layer, rather than learning both lexical tokens and how to use them in-context.

Both case agreement and syntactic error showed moderate to very strong relationships with pronoun consistency (CA Spearman $\rho = .841$, SE Spearman $\rho = -.644$,). T_{PTP} + M_{Lex} resulted in the highest case agreement (32.5%), as shown in Table 1. Interestingly, even when trained on more data and using both training regimes, syntactic errors did not decrease as significantly with T_{Orig} as with T_{PTP}, suggesting that PTP helps an LLM improve its grammatical usage of neopronouns. Additionally, M_{Lex} for 70M resulted in an **18.8% reduction in training time** compared to M_{Full}, presenting a more resource-efficient and eco-friendly alternative to standard full fine-tuning (more results in Appendix Table 4). Regarding binary pronouns, the effects of lexical fine-tuning varied for models of this size, with some showing temporary increases in consistency and others improving with full fine-tuning. This may be related to nuances in lexically finetuning existing versus new pronouns. Binary pronouns, having already converged in latent lexical space, may subtly shift during finetuning. However, neopronouns are not initialized from an initial space of semantic relevance, thereby requiring the LLM to learn them from scratch. Consequently, while the LLM learns these new neopronouns, the already trained binary pronouns may be impacted. These trends extended to case agreement and syntactic error, though the variations in binary pronoun values were not statistically different from lexical finetuning without PTP.

5.3 Experiment 3: Does a bigger LLM always mean better pronoun consistency?

In this experiment, our questions consider scale: (1) Does our methodology work for model sizes beyond 70M? (2) Does an increase in model capacity always mean improved pronoun consistency? And if not, (3) at what point do we maximize improvements with PTP? To answer these, we run experiments encompassing the previous experiments across 70M, 160M, 410M, and 1.4B parameter models. Each model is finetuned and evaluated in the same way as prior experiments, using only a 10% resource split in order to best reflect realistic low-resource scenarios with neopronouns.

Results Our results highlight the effectiveness of lexical training for PTP across different model sizes, with smaller models achieving performance boosts comparable to models more than twice their size. Figure 6 demonstrates that, across all model sizes, $T_{\text{PTP}} + M_{\text{Lex}}$ consistently improved pronoun consistency over T_{Orig} across finetuning regimes. Across both PTP and lexical finetuning respectively, we find these models maintain high quality generations, as shown in Table 2. We consistently

Table 2: Generations from Pythia-410M from the prompt: *Skyler's pronouns are xe/xem/xir/xirself.*

Generations	Text
T_{Orig} + M_{Baseline}	...Xe needs to go to the bookstore soon because it's not in the store. Xir is the same as Xirself.
T_{Orig} + M_{Full}	...Xe needs to go to the bookstore soon because xe's too busy to be on the air. Xe's also a good friend of xir.
T_{PTP} + M_{Full}	...Xe needs to go to the bookstore soon because xe's going to be writing a book about how to say xir name. Xe also has a book in the works called "the art of being a writer.
T_{PTP} + M_{Lex}	...Xe is a huge fan of the book "the secret life of the apes" by john mccarthy.

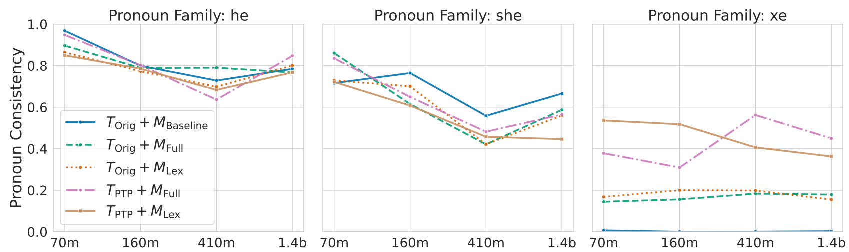

Figure 6: Misgendering Rates across model sizes. Highest consistency for neopronouns seen in models using PTP.

observed case agreement improvements for certain binary pronouns when M_{Lex} was added (see Appendix Table 8). Between M_{Lex} and M_{Full}, PTP was most effective with lexical finetuning in models smaller than 410M parameters, while finetuning proved more beneficial for larger models. The highest pronoun consistency for *xe* was achieved using PTP with a 410M parameter model (58%) using full finetuning, followed by a 160M parameter model using only lexical finetuning (55%). While the 410M model did show lower syntactic error (27%, vs 35% for 160M) and a higher average case agreement (48% vs. 40%), these results indicate that high levels of pronoun consistency seen in large models can also be attained in smaller models given appropriate mitigation strategies.

6 Conclusion

In this work, we demonstrate how LLM misgendering is influenced by low-resource BPE tokenization. We find that We demonstrate this empirically, finding that LLM misgendering closely linked its inability to adhere to neopronoun morphosyntax. With this knowledge, we propose a mitigation procedure called *pronoun tokenization parity*, designed to preserve neopronoun functional structure through special tokens. We find that LLMs finetuned with PTP reduces neopronoun misgendering over traditional finetuning settings without PTP. Likewise, exploiting pre-existing English grammatical knowledge with PTP also achieves effective mitigation in a cost-effective manner. As BPE tokenization is just one of many subword tokenization algorithms, our work lays groundwork for exploring misgendering in LLMs across these schemes. Future work may also study misgendering due to tokenization within a multilingual setting.

Limitations and Broader Impacts

As neopronouns continue to surface and be adopted, we highlight the importance of considering how each pronoun family operates within its own language. Therefore, we show this as an end-to-end example for one pronoun family, *xe* with respect to English. Future work should also consider the how pronoun family operates within the LLM as well. Furthermore, adding other metrics from existing bias benchmarks may complement our study. We mostly rely on quantitative metrics grounded in English grammar rules to assess the quality of mitigations. While an in-depth human evaluation may also supplement this work, we qualitatively evaluate resulting text generations from our approaches (please see Appendix Table 2).

We also emphasize the importance of transparent stakeholder discourse in selecting an approach that balances pronoun consistency, error rates, and case agreement. For instance, if choosing to address historical disparities for minority groups, stakeholders may choose to prioritize their improvement while specifying a error tolerance for dominant groups rather than solely aiming for equal or improved performance across majority groups.

References

Ali Araabi, Christof Monz, and Vlad Niculae. How effective is byte pair encoding for out-of-vocabulary words in neural machine translation? In *Conference of the Association for Machine Translation in the Americas*, 2022. URL `https://api.semanticscholar.org/CorpusID:251468147`.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Bilingual lexicon induction through unsupervised machine translation. *ArXiv*, abs/1907.10761, 2019a. URL `https://api.semanticscholar.org/CorpusID:196187034`.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Annual Meeting of the Association for Computational Linguistics*, 2019b. URL `https://api.semanticscholar.org/CorpusID:204901567`.

Xingce Bao and Qianqian Qiao. Transfer learning from pre-trained bert for pronoun resolution. In *Proceedings of the first workshop on gender bias in natural language processing*, pages 82–88, 2019.

Stella Rose Biderman, Hailey Schoelkopf, Quentin G. Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. *ArXiv*, abs/2304.01373, 2023. URL `https://api.semanticscholar.org/CorpusID:257921893`.

Kaj Bostrom and Greg Durrett. Byte pair encoding is suboptimal for language model pretraining. In *Findings*, 2020. URL `https://api.semanticscholar.org/CorpusID:215416175`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

František Dařena and Martin Suss. Quality of word vectors and its impact on named entity recognition in czech. *European Journal of Business Science and Technology*, 2020. URL `https://api.semanticscholar.org/CorpusID:231686297`.

Wietse de Vries and Malvina Nissim. As good as new. how to successfully recycle english gpt-2 to make models for other languages. *ArXiv*, abs/2012.05628, 2020. URL `https://api.semanticscholar.org/CorpusID:228083868`.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.150. URL `https://aclanthology.org/2021.emnlp-main.150`.

Miguel Domingo, Mercedes García-Martínez, Alexandre Helle, Francisco Casacuberta, and Manuel Herranz. How much does tokenization affect neural machine translation? In *Conference on Intelligent Text Processing and Computational Linguistics*, 2018. URL `https://api.semanticscholar.org/CorpusID:56517054`.

Virginia K. Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. *ArXiv*, abs/2306.15087, 2023. URL `https://api.semanticscholar.org/CorpusID:259262064`.

M.D. Fortescue. *Historical Linguistics 2003: Selected Papers from the 16th International Conference on Historical Linguistics, Copenhagen, 11-15 August 2003*. Amsterdam Studies in the Theory and History of Linguistic Science: 4. J. Benjamins Pub., 2005. ISBN 978-1-58811-586-7.

B.A. Garner. *The Chicago Guide to Grammar, Usage, and Punctuation*. Chicago Guides to Writing, Editing, and Publishing. University of Chicago Press, 2016. ISBN 978-0-226-19129-4.

Gender Census. 2023 gender census, 2023. URL https://www.gendercensus.com/results/2023-worldwide/#pronouns. Accessed: September 14, 2023.

Ximena Gutierrez-Vasques, Christian Bentz, O. A. Sozinova, and Tanja Samardvzic. From characters to words: the turning point of bpe merges. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2021. URL https://api.semanticscholar.org/CorpusID:233189533.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020. doi: 10.5281/zenodo.1212303.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. Misgendered: Limits of large language models in understanding pronouns. *arXiv preprint arXiv:2306.03950*, 2023.

Matthias Huck, Viktor Hangya, and Alexander M. Fraser. Better oov translation with bilingual terminology mining. In *Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:196193105.

Anne Lauscher, Archie Crowley, and Dirk Hovy. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. *arXiv preprint arXiv:2202.11923*, 2022.

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and A. G. Galstyan. Man is to person as woman is to location: Measuring gender bias in named entity recognition. *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, 2019. URL https://api.semanticscholar.org/CorpusID:204851964.

Ehm Hjorth Miltersen. Nounself pronouns: 3rd person personal pronouns as identity expression. *Journal of Language Works-Sprogvidenskabeligt Studentertidsskrift*, 1(1):37–62, 2016.

Debora Nozza, Federico Bianchi, Anne Lauscher, Dirk Hovy, et al. Measuring harmful sentence completion in language models for lgbtqia+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics, 2022.

Organizers of QueerInAI, Nathaniel Dennler, Anaelia Ovalle, Ashwin Singh, Luca Soldaini, Arjun Subramonian, Huy Tu, William Agnew, Avijit Ghosh, Kyra Yee, Irene Font Peradejordi, Zeerak Talat, Mayra Russo, and Jessica de Jesus de Pinho Pinhal. Bound by the bounty: Collaboratively shaping evaluation processes for queer ai harms. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023. URL https://api.semanticscholar.org/CorpusID:259991807.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1246–1266, 2023.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276, 2020. URL https://api.semanticscholar.org/CorpusID:228375396.

Penny Eckert and Ivan A. Sag. Morphology, 2011. URL https://web.stanford.edu/class/linguist1/Slides/morph2-slides.pdf. Accessed: 2023.

Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. Mimicking word embeddings using subword rnns. *ArXiv*, abs/1707.06961, 2017. URL https://api.semanticscholar.org/CorpusID:10361075.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*, 2018.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.243. URL https://aclanthology.org/2021.acl-long.243.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009.

Karolina Stanczak and Isabelle Augenstein. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*, 2021.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*, 2019.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

Zeerak Talat and Anne Lauscher. Back to the future: On potential histories in nlp. *ArXiv*, abs/2210.06245, 2022. URL https://api.semanticscholar.org/CorpusID:252846634.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, and Dong Yu. Improving pre-trained multilingual model with vocabulary expansion. *ArXiv*, abs/1909.12440, 2019. URL https://api.semanticscholar.org/CorpusID:203586495.

Xiao Wang, Shihan Dou, Li Xiong, Yicheng Zou, Qi Zhang, Tao Gui, Liang Qiao, Zhanzhan Cheng, and Xuanjing Huang. Miner: Improving out-of-vocabulary named entity recognition from an information theoretic perspective. In *Annual Meeting of the Association for Computational Linguistics*, 2022. URL https://api.semanticscholar.org/CorpusID:248085423.

Alfan Farizki Wicaksono and Ayu Purwarianti. Hmm based part-of-speech tagger f or bahasa indonesia. 2010. URL https://api.semanticscholar.org/CorpusID:63251684.

Shaked Yehezkel and Yuval Pinter. Incorporating context into subword vocabularies. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2022. URL https://api.semanticscholar.org/CorpusID:252872957.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *North American Chapter of the Association for Computational Linguistics*, 2018. URL https://api.semanticscholar.org/CorpusID:4952494.

A Appendix

A.1 Details on How to Reproduce Pronoun Tokenization Parity (PTP)

We provide details on how to reproduce PTP in Algorithm 1.

Algorithm 1 Pronoun Tokenization Parity (PTP)

Input 1: LLM model
Input 2: LLM model's BPE tokenizer
Input 3: Defined list of neopronouns for PTP
Input 4: Dataset augmented with neopronouns
Method: Add special tokens for each neopronoun. Be sure to explicitly add 'Ġ' to the beginning of each token to indicate that it is a full, non-subword token space before the word, otherwise this will lead to incorrect model behavior, since a lack of 'Ġ' in BPE tokenization indicates a subword token.
Check: Check the tokenizer is working properly by checking the tokenized neopronoun, ensuring that you see 'Ġ' in its token. For example, tokenizing *xe* should result in ['Ġxe'] not ['Ġ', xe']. The latter will cause the LLM to incorrectly associate a space character with a neopronoun. This can be tested by checking next word transition probabilities from the space character.
Resize the LLM token embeddings to match vocabulary of tokenizer. Here is example code to do this with a model and tokenizer from `HuggingFace Transformers` Package [5].

```
#declare neopronoun tokens
arr_tokens = [
    'Ġxe', 'ĠXe',
    'Ġxem', 'ĠXem',
    'Ġxir', 'ĠXir',
    'Ġxirs', 'ĠXirs'
]

# add new tokens to the tokenizer, t
token_dict = {
'additional_special_tokens': arr_tokens
}
t.add_special_tokens(token_dict)

# update model, m, accordingly
m.resize_token_embeddings(len(tokenizer))
```

if `Lexical Finetuning` **then**
    Freeze all parameters besides the word token embeddings. Then proceed to finetune this lexical layer.
else
    Proceed with standard full finetuning
Return Finetuned model, new PTP tokenizer
Evaluate using extended `MISGENDERED` framework

A.2 Measuring Fertility

We measure "fertility" as defined by Rust et al. [2021]. We define binary pronouns as pronouns across all cases for *he* and *she*. We define the neopronoun group as all pronoun cases for common neopronouns *xe*, *ey*, and *fae*. [6] 10k most common words are determined by n-gram frequency analysis of the Google's Trillion Word Corpus. [7]

[6] https://nonbinary.wiki/wiki/English_neutral_pronouns
[7] https://github.com/first20hours/google-10000-english/blob/master/google-10000-english-usa-no-swears-short.txt

A.3 Embedding Initialization

Upon adding a new token and creating a new E_{PTP}, embeddings are set to default random initialization behavior in an LLM. Being that neopronouns and binary pronouns follow the same grammar rules in English, we also investigate leveraging *existing* grammatical knowledge learned by the LLM to help bootstrap the model's ability to learn to use neopronouns better. Establishing a direct mapping between binary and neopronouns across their various forms, we average the neopronoun embedding with its corresponding binary pronoun embedding for each case. This approach resembles the use of a bilingual lexicon to facilitate vocabulary alignment Artetxe et al. [2019a].

We adopt the method of taking the mean across binary pronouns for two key reasons: to leverage the LLM's syntactic knowledge related to singular pronouns used similarly to *xe* in sentences and to accommodate individuals who use neopronouns and may have historical associations with binary pronouns. This is denoted in the tables from Section A.6 as PTP-B. For future work, we encourage further exploration of methods to bootstrap these embeddings.

A.4 Model Finetuning Details

A.4.1 Experiment 1 - Full Finetuning

We use the *deduped* versions of Pythia, which trained on the Pile after the dataset had been globally deduplicated. We confirm that our research is in line with Pythia's intended use: Given their Apache 2.0 license, we may finetune or adapt these models.

Before tokenization, text is chunked with a 256 window size, resulting in 386,267 rows before any neopronoun augmentation. We conduct fine-tuning with an 80/10/10 train, validation, and test split. Each model adheres to Pythia suite configurations, including an embedding size of 512 and a vocabulary size of 50,284 (50,277 without PTP). Fine-tuning is done for five epochs with a batch size of 128, a learning rate of 10^{-4}, and early stopping based on cross-entropy loss in the validation set with a patience of 2. To expedite model training, all models undergo fine-tuning using FP16 mixed precision and 2 gradient accumulation steps.

A.4.2 Experiment 2 - Lexical Training

We follow the setup from the previous experiment, but only slightly increase the learning rate to 1 x 10^3 in order to encourage more rapid adaptation to the new vocabulary.

A.4.3 Hardware Setup

We perform all our experiments with 8 NVIDIA A100s with 40 GiB vRAM.

Table 3: Average GPU Hours For Full Finetuning

Model Size	Hours
70M	0.65
160M	0.74
410M	1.2
1.4B	1.7

Table 4: Δ Training runtime upon switching to lexical finetuning.

Model Size	Training Time Reduction (%)
70M	18.8
160M	21.1
410M	16.5
1.4B	21.5

Table 5: Model Parameters (P), Available on HuggingFace.

Model Size	# P	# Non-Embedding P
70M	70,426,624	18,915,328
160M	162,322,944	85,056,000
410M	405,334,016	302,311,424
1.4B	1,414,647,808	1,208,602,624

A.5 Templates additions to MISGENDERED

To mimic real world pronoun declarations, each declaration is started with nominative, accusative, pronominal possessive, and reflexive pronouns.

Table 6 reflects selected additions from the TANGO dataset. Det represents the determiner position one may replace with ones like *the, a, these, those*. Gen-dep, Gen-indep, reflex, nom are all pronoun cases.

Table 6: Template additions to MISGENDERED Hossain et al. [2023]

Case	Template
Nominative	before {name} came out, {det} {nom} identified differently. {name} is an indian filmmaker known for directing a movie, however {det} {nom} also directs television. {name} is a musician and dancer and {det} {nom} currently lives somewhere nearby.
Accusative	upon recognizing {name}, the fan asked {det} {acc} for an autograph. {name} owns the car so be sure to give the keys back to {det} {acc} afterwards. {name} needs new shoes so I can give my other shoes to {det} {acc}.
Pronominal Possessive	{name} was a participant in {det} {gen-indep} own musical. {name} is an author of children’s fantasy, best known for {det} {gen-indep} book that won several awards. {name} is a singer-songwriter-producer based in boston and {det} {gen-indep} debut album was released in march.
Predicative Possessive	{name} was working in this area, so the phone is probably {det} {gen-dep}. the story of {name}’s resilience and triumph over adversity is {det} {gen-dep} to own. name} said to me that the larger slice of pizza was mine, and the smaller one was {det} {gen-dep}.
Reflexive	{name} will read the book by {det} {reflex}. {name} needs to be by {det} {reflex} sometimes. {name} often works alone by {det} {reflex}.

A.6 Ablations

Table 7 provides results across all data splits for the 70M model. Table 8 provides results across model sizes for the 10% data resource ablation, so as to best mimic real-world low-resource circumstances.

Table 7: 70M Model Results Across Data Splits

(a) Data Split= 10

Model	Pronoun Consistency (↑) He	She	Xe	Case Agreement (↑) He	She	Xe	Error (↓) He	She	Xe
$T_{\text{Orig}} + M_{\text{Baseline}}$	0.968	0.716	0.007	0.677	0.607	0.200	0.238	0.169	0.850
$T_{\text{Orig}} + M_{\text{Full}}$	0.896	**0.861**	0.145	0.685	0.597	0.200	**0.239**	**0.168**	0.895
$T_{\text{Orig}} + M_{\text{Lex}}$	0.865	0.729	0.168	0.617	0.572	0.200	0.290	0.232	0.654
$T_{\text{PTP}} + M_{\text{Full}}$	0.948	0.835	0.378	**0.695**	0.611	0.294	0.278	0.210	0.270
$T_{\text{PTP-B}} + M_{\text{Full}}$	**0.962**	0.807	0.244	0.689	**0.612**	0.231	0.283	0.206	**0.259**
$T_{\text{PTP}} + M_{\text{Lex}}$	0.850	0.722	**0.536**	0.632	0.591	0.325	0.258	0.218	0.348
$T_{\text{PTP-B}} + M_{\text{Lex}}$	0.833	0.743	0.430	0.654	0.596	**0.332**	0.243	0.202	0.321

(b) Data Split= 20

Model	Consistency (↑) He	She	Xe	Case (↑) He	She	Xe	Error (↓) He	She	Xe
$T_{\text{Orig}} + M_{\text{Baseline}}$	0.968	0.715	0.007	0.677	0.716	0.200	0.238	0.169	0.850
$T_{\text{Orig}} + M_{\text{Full}}$	0.932	0.814	0.137	0.666	**0.853**	0.200	0.271	**0.180**	0.873
$T_{\text{Orig}} + M_{\text{Lex}}$	0.861	0.733	0.181	0.638	0.573	0.200	0.276	0.199	0.675
$T_{\text{PTP}} + M_{\text{Full}}$	**0.965**	**0.886**	0.359	**0.719**	0.597	**0.364**	**0.253**	0.192	**0.338**
$T_{\text{PTP-B}} + M_{\text{Full}}$	0.953	0.873	0.185	0.701	0.598	0.325	0.269	0.195	0.340
$T_{\text{PTP}} + M_{\text{Lex}}$	0.822	0.709	**0.480**	0.605	0.586	0.322	0.301	0.237	0.349
$T_{\text{PTP-B}} + M_{\text{Lex}}$	0.832	0.701	0.324	0.639	0.583	0.328	0.301	0.226	0.340

(c) Data Split= 30

Model	Consistency (↑) He	She	Xe	Case (↑) He	She	Xe	Error (↓) He	She	Xe
$T_{\text{Orig}} + M_{\text{Baseline}}$	0.968	0.716	0.007	0.677	0.607	0.200	0.238	0.169	0.850
$T_{\text{Orig}} + M_{\text{Full}}$	0.913	0.852	0.138	0.689	**0.614**	0.200	0.246	**0.191**	0.870
$T_{\text{Orig}} + M_{\text{Lex}}$	0.801	0.647	0.185	0.627	0.574	0.200	0.292	0.243	0.666
$T_{\text{PTP}} + M_{\text{Full}}$	**0.958**	**0.877**	0.324	**0.724**	0.603	0.357	**0.232**	0.197	0.344
$T_{\text{PTP-B}} + M_{\text{Full}}$	0.909	0.844	0.126	0.721	0.601	0.288	0.255	**0.191**	0.260
$T_{\text{PTP}} + M_{\text{Lex}}$	0.812	0.620	0.455	0.638	0.566	**0.348**	0.268	0.209	0.314
$T_{\text{PTP-B}} + M_{\text{Lex}}$	0.849	0.693	**0.483**	0.648	0.573	0.317	0.277	0.210	0.333

(d) Data Split= 40

Model	Consistency (↑) He	She	Xe	Case (↑) He	She	Xe	Error (↓) He	She	Xe
$T_{\text{Orig}} + M_{\text{Baseline}}$	0.968	0.716	0.007	0.677	0.607	0.200	0.238	0.169	0.850
$T_{\text{Orig}}+ M_{\text{Full}}$	0.924	0.793	0.099	0.670	**0.607**	0.200	0.254	**0.196**	0.855
$T_{\text{Orig}}+ M_{\text{Lex}}$	0.822	0.652	0.182	0.637	0.564	0.200	0.311	0.226	0.682
$T_{\text{PTP}}+ M_{\text{Full}}$	0.960	**0.864**	0.268	0.669	0.588	0.321	0.251	0.200	0.339
$T_{\text{PTP-B}}+ M_{\text{Full}}$	**0.967**	0.862	0.117	**0.677**	0.594	0.307	**0.234**	0.203	0.339
$T_{\text{PTP}}+ M_{\text{Lex}}$	0.853	0.615	**0.484**	0.644	0.572	0.331	0.290	0.216	0.332
$T_{\text{PTP-B}}+ M_{\text{Lex}}$	0.849	0.620	0.414	0.636	0.573	**0.337**	0.289	0.229	**0.331**

(e) Data Split= 50

Model	Consistency (↑) He	She	Xe	Case (↑) He	She	Xe	Error (↓) He	She	Xe
$T_{\text{Orig}} + M_{\text{Baseline}}$	0.968	0.716	0.007	0.677	0.607	0.200	0.238	0.169	0.850
$T_{\text{Orig}}+ M_{\text{Full}}$	0.934	0.852	0.141	**0.664**	**0.613**	0.200	0.260	0.198	0.861
$T_{\text{Orig}}+ M_{\text{Lex}}$	0.834	0.651	0.185	0.610	0.567	0.200	0.287	**0.190**	0.712
$T_{\text{PTP}}+ M_{\text{Full}}$	**0.960**	**0.889**	0.267	0.649	0.604	0.335	**0.244**	0.217	0.359
$T_{\text{PTP-B}}+ M_{\text{Full}}$	0.950	0.872	0.161	0.626	0.590	0.320	0.295	0.216	0.379
$T_{\text{PTP}}+ M_{\text{Lex}}$	0.775	0.582	**0.584**	0.625	0.572	**0.366**	0.291	0.197	0.315
$T_{\text{PTP-B}} + M_{\text{Lex}}$	0.815	0.644	0.493	0.637	0.557	0.354	0.268	0.224	**0.305**

Table 8: Model Size Comparisons at Data Split=10

(a) 160M Parameter Model Results

Model	Consistency (↑)			Case (↑)			Error (↓)		
	He	She	Xe	He	She	Xe	He	She	Xe
$T_{\text{Orig}} + M_{\text{Baseline}}$	0.799	0.765	0.000	0.745	0.644	0.199	0.087	0.065	0.954
$T_{\text{Orig}} + M_{\text{Full}}$	0.789	0.615	0.156	0.716	0.609	0.201	0.194	0.203	0.692
$T_{\text{Orig}} + M_{\text{Lex}}$	0.773	**0.701**	0.200	0.703	0.593	0.200	0.202	0.169	0.783
$T_{\text{PTP}} + M_{\text{Full}}$	0.802	0.649	0.309	**0.740**	**0.606**	0.289	0.220	0.182	**0.147**
$T_{\text{PTP-B}} + M_{\text{Full}}$	0.791	0.658	0.097	0.667	0.599	0.244	0.209	0.210	0.253
$T_{\text{PTP}} + M_{\text{Lex}}$	0.785	0.608	0.518	0.729	0.599	**0.398**	0.191	**0.148**	0.314
$T_{\text{PTP-B}} + M_{\text{Lex}}$	**0.812**	0.605	**0.536**	0.734	0.602	0.336	**0.175**	0.164	0.251

(b) 410m Parameter Model Results

Model	Consistency (↑)			Case (↑)			Error (↓)		
	He	She	Xe	He	She	Xe	He	She	Xe
$T_{\text{Orig}} + M_{\text{Baseline}}$	0.728	0.559	0.001	0.767	0.663	0.200	**0.042**	0.035	**0.898**
$T_{\text{Orig}} + M_{\text{Full}}$	0.790	0.421	0.184	0.744	0.617	0.201	0.128	0.198	0.566
$T_{\text{Orig}} + M_{\text{Lex}}$	0.698	0.421	0.198	0.765	0.653	0.210	0.170	**0.118**	0.545
$T_{\text{PTP}} + M_{\text{Full}}$	0.636	0.482	**0.562**	0.737	0.620	0.406	0.144	0.150	0.207
$T_{\text{PTP-B}} + M_{\text{Full}}$	**0.823**	**0.488**	0.190	0.732	0.612	0.417	**0.115**	0.174	0.262
$T_{\text{PTP}} + M_{\text{Lex}}$	0.683	0.457	0.406	0.765	0.641	0.478	0.167	0.134	0.136
$T_{\text{PTP-B}} + M_{\text{Lex}}$	0.694	0.356	0.492	**0.774**	**0.655**	**0.487**	0.177	0.174	**0.122**

(c) 1.4B Parameter Model Results

Model	Consistency (↑)			Case (↑)			Error (↓)		
	He	She	Xe	He	She	Xe	He	She	Xe
$T_{\text{Orig}} + M_{\text{Baseline}}$	0.785	0.665	0.003	0.782	0.714	0.200	0.037	0.034	0.928
$T_{\text{Orig}} + M_{\text{Full}}$	0.767	**0.587**	0.179	0.750	0.628	0.201	0.247	0.246	0.363
$T_{\text{Orig}} + M_{\text{Lex}}$	0.801	0.560	0.155	0.793	**0.646**	0.200	0.166	0.355	0.552
$T_{\text{PTP}} + M_{\text{Full}}$	**0.847**	0.565	**0.450**	0.730	0.661	0.450	0.241	**0.180**	0.201
$T_{\text{PTP-B}} + M_{\text{Full}}$	0.719	0.539	0.357	0.734	0.629	0.447	0.191	0.223	**0.185**
$T_{\text{PTP}} + M_{\text{Lex}}$	0.768	0.446	0.363	**0.818**	0.623	0.383	**0.121**	0.246	0.193
$T_{\text{PTP-B}} + M_{\text{Lex}}$	0.797	0.578	0.356	0.809	0.627	**0.482**	0.201	0.267	0.272