The SlayQA Benchmark of Social Reasoning: Testing Gender-inclusive Generalization with Neopronouns

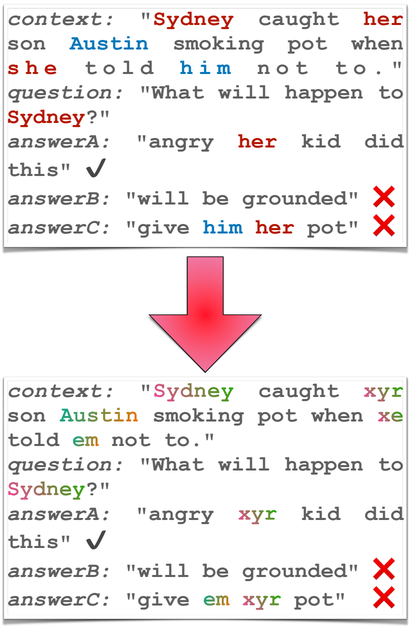
Bastian Bunzeck and **Sina Zarrieß**
Computational Linguistics, Department of Linguistics
Bielefeld University, Germany
{bastian.bunzeck, sina.zarriess}@uni-bielefeld.de

Abstract

We introduce SlayQA, a novel benchmark data set designed to evaluate language models' ability to handle gender-inclusive language, specifically the use of neopronouns, in a question-answering setting. Derived from the Social IQa data set, SlayQA modifies context-question-answer triples to include gender-neutral pronouns, creating a significant linguistic distribution shift in comparison to common pretraining corpora like C4 or Dolma. Our results show that state-of-the-art language models struggle with the challenge, exhibiting small, but noticeable performance drops when answering question containing neopronouns compared to those without.

1 Introduction

Currently, the recognition of the importance of inclusivity and representation in NLP is growing (Sun et al., 2019; Stanczak and Augenstein, 2021; Lauscher et al., 2022). Traditional data sets often reflect and perpetuate binary gender norms, which can marginalize non-binary and gender non-conforming individuals and cause harm (Ansara and Hegarty, 2013). This lack of inclusivity highlights a critical need for resources that better represent the full spectrum of gender identities. One aspect of gender-inclusive language use that is gaining more and more acceptance is the usage of neopronouns like *xe/xyr* or *ze/zir*. Neopronouns are novel pronouns that people who do not identify themselves as belonging to the polar extremes of the gender spectrum can choose to use for reference to themselves instead of the classical, gendered pronouns: *he/him* and *she/her*. Current benchmarks that assess this kind of linguistic inclusivity mostly focus on the generation of correct neopronouns in context or similar tasks (Ovalle et al., 2023; Hossain et al., 2023). On the other hand, resources that simply implement established LM benchmarks in a more inclusive way are rare.



context: "Sydney caught her son Austin smoking pot when she told him not to."
question: "What will happen to Sydney?"
answerA: "angry her kid did this" ✓
answerB: "will be grounded" ✗
answerC: "give him her pot" ✗

context: "Sydney caught xyr son Austin smoking pot when xe told em not to."
question: "What will happen to Sydney?"
answerA: "angry xyr kid did this" ✓
answerB: "will be grounded" ✗
answerC: "give em xyr pot" ✗

Figure 1: Visualization of the conversion process that turns Social IQa data into SlayQA data

To address this gap, we present SlayQA: **S**ocial **l**inguistics **a**nalytics **y**ielding **Q**ueer **A**gents, a novel benchmark set derived from the existing Social IQa (SIQa) data set (Sap et al., 2019b). It contains a situation description (the context), social reasoning questions and three prospective answers, where all context-question-answer pairs include at least two acts of pronoun-based reference and gender-neutral pronouns. Because SlayQA systematically replaces established, gendered pronouns with gender-affirming neopronouns, it is a more inclusive data set that better reflects the diversity of human identities. Although neo-pronouns are commonly rated less grammatical than their established counterparts (Hekanaho, 2021; Rose et al., 2023), they are beginning to be adopted in various social circles. Here, the key word is *beginning* – they are still infrequent in discourse and and also in common pretraining corpora like C4 (Raffel et al., 2020), Dolma (Soldaini et al., 2024), and

Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP, pages 42–53
November 16, 2024 ©2024 Association for Computational Linguistics

RedPajama-Data-1T (Together Computer, 2023). As such, our pronoun-altered benchmark marks a significant pronoun distribution shift in comparison to these pretraining corpora. Consequentially, SlayQA helps to assess how well language models are able to generalize to novel linguistic structures.

2 Related work

Dev et al. (2021) conduct a survey on harms involving gender-neutral speech and neopronouns in general purpose NLP systems. Replies included, for example, the non-detection of hate-speech or automatic educational assessments marking gender-inclusive language as wrong. Furthermore they show that from skewed training data, bias in word embeddings arises. For GLoVe embeddings (Pennington et al., 2014), gendered pronouns appear in close proximity in their vector space, whereas neopronouns hardly cluster with them. Similarly, in the original BERT model (Devlin et al., 2019), neopronouns are out-of-vocabulary items. In a more applied setting, Lauscher et al. (2023) show that machine translation systems are able to translate gendered pronouns well, but not neopronouns. They are either plainly copied or the included agents are misgendered. Furthermore, for sentences with gender-neutral or neopronouns, overall translation quality (e.g. syntactic, semantic) diminishes.

The most prominent NLP benchmarks for gender-inclusivity are TANGO (Ovalle et al., 2023) and MISGENDERED (Hossain et al., 2023). While TANGO contains sentences with names and neopronouns to be completed by generative models, MISGENDERED also includes an explicit statement of the agents' preferred pronouns. Their goals are therefore almost identical: to assess whether language models can correctly produce text with neopronouns when prompted with them. Evaluations on these data sets show similar results: errors rarely occur with gendered pronouns, but correct continuation scores drop with the gender-neutral singular *they*. Worst scores (accuracy below 10%) are found for neopronouns. A possible explanation can be found in Ovalle et al. (2024), who show that the BPE algorithm (Gage, 1994) commonly used in state-of-the-art LLMs dissects neopronouns into smaller parts, which never happens to established pronouns. This is caused by data scarcity – common pre-training data sets lack examples of neopronouns in use. As the BPE algorithm leaves lexical tokens intact if and only if they occur with a high enough frequency, neopronouns are usually split.

	C4	Dolma
he	144.202.977	965.297.366
she	92.421.725	544.245.250
they	260.126.090	1.705.400.768
thon	872.654	992.499
e	213.797.769	240.457.628
ae	3.910.812	4.135.288
co	83.935.707	199.206.147
vi	10.139.390	12.534.070
xe	1.148.568	2.134.212
ey	869.765	1.691.904
ze	1.618.896	1.793.116

Table 1: Frequencies for established subject pronouns and subject neopronouns in C4 and Dolma

Current data sets that are used to measure the question answering abilities of NLP systems are not concerned with gender-inclusivity. While Rogers et al. (2023) present a large taxonomy including many different kinds of tasks, domains and data formats, 'fairness' seems to be only an afterthought in contemporary QA evaluation, e.g. by only referring to the inclusion of multilingual data.

3 (Neo)pronouns in pre-training corpora and evaluation data sets

3.1 (Neo)pronouns in C4 and Dolma

We argue that our benchmark introduces a significant distribution shift between the pretraining corpora and the evaluation data with regard to pronouns. To assess this proposed distribution shift, we determine the frequencies of established and neopronouns in these corpora through the n-gram lookup function of *What's In My Big Data?* (WIMBD) (Elazar et al., 2024) – if the neopronouns occur less frequently in pretraining corpora than established pronouns, then our neopronoun benchmark introduces a drastic distribution shift in its pronoun distribution compared to these corpora.

We adapt our list of pronouns from the seminal study by Hossain et al. (2023). For the sake of brevity, we do not include further gender-affirming pronoun variations like nounself, emojiself, numberself or nameself pronouns (Lauscher et al., 2022).

We search C4 (Raffel et al., 2020) and Dolma (Soldaini et al., 2024) for the subject, object, possessive (pronoun and determiner) and reflexive

	context	question	answerA	answerB	answerC
he	10.238	146	2.020	2.066	2.060
she	13.291	221	2.759	2.716	2.715
they	14.178	150	2.996	2.993	3.007
thon	0	0	0	0	0
e	1	0	0	0	0
ae	0	0	0	0	0
co	0	0	0	0	0
vi	0	0	0	0	0
xe	0	0	0	0	0
ey	0	0	0	0	0
ze	0	0	0	0	0

Table 2: Token frequencies for morphological paradigms of gendered and gender-neutral pronouns in Social IQa data

forms of the neopronouns from Hossain et al. (2023), as we evaluate models trained on these corpora. While we also evaluate models trained on RedPajama-Data-1T (Together Computer, 2023), no n-gram frequencies for this data set are available through WIMBD. C4 is based on CommonCrawl web dumps that were then cleaned, filtered and deduplicated to certain degrees. RedPajama-Data-1T and Dolma also contain considerable portions of CommonCrawl enriched with data from diverse sources, such as GitHub code, Reddit posts, academic papers from SemanticScholar and Arxiv, etc., which were then also cleaned, filtered and deduplicated. They were explicitly created as open data sets that mirror the data that commercial/closed models like Anthropic's Claude, OpenAI's ChatGPT or Meta's Llama models are trained on. Therefore, they can be seen as somewhat exemplary for the data that commercial models are trained on, and the overall frequency distributions found in them should be generally similar to those in not publicly available pre-training corpora.

The frequencies for subject pronouns in both corpora are found in Table 1, all other results are listed in Appendix A. In comparison to the established pronouns, neopronouns occur with reduced frequencies. The neopronouns *e* and *co* are exceptions, but as *e* is a highly frequent letter in the English language and *co* also serves as a productive morpheme, it is reasonable to assume that the vast majority of these instances are not representative of pronoun usage. The neopronouns that do not constitute such widely used building blocks of ordinary English, e.g. *thon*, *xe* or *ey*, occur much less throughout the training data. For example, *he* occurs one thousand times more than *thon* in Dolma. These distributions are stable across all morphological forms (see Tables 5, 6, 7 and 8). Although the possesive and reflexive pronouns are overall less frequent, all forms are still found across all training corpora. The only exception is the reflexive *virself*, which is completely absent from the C4 data. Yet, the presence of all other forms in the data, and especially the presence of the reflexives, which should not be accidental n-gram matches, confirms that pronoun use of these neopronouns is indeed included in these pre-training data sets, just to a much lesser degree than the usage of established pronouns.

Although no n-gram frequency results are available for RedPajama-Data-1T, we assume that the underlying distribution should be mostly equal to the two examined corpora – all three corpora are mainly based on CommonCrawl web dumps, so it is reasonable to expect a large lexical overlap between them.

A final indicator for the different pronoun distributions can be found in Table 9, where we show the number of sub-word tokens which the different grammatical forms of our investigated (neo)pronouns are split into by the tokenizers of our tested models. Here, we find generally higher numbers for the neopronouns, especially for the reflexive forms. Because the standard BPE tokenization algorithm keeps highly frequent forms intact as one token, this split of lexical words into several sub-words is another display of their infrequency compared to established pronouns.

3.2 (Neo)pronouns in SIQa

While the mentioned pre-training corpora contain very little neopronouns, the numbers are even more extreme for the SIQa data set (Sap et al., 2019b). The original SIQa contains 37.588 triples of context, question and three prospective answers. It is based on the Atomic data set (Sap et al., 2019a), which contains commonsense if-then statements for machine learning. These were then manually rewritten into context, question and right answer triples. False answers were added manually and by sampling randomly from correct answers to different questions. As gender fairness was not a concern in the compilation of this data set, the distributions of gendered pronouns and gender-neutral neopronouns deviate strongly. Table 2 shows the absolute token counts (aggregating over the complete morphological paradigms) for SIQa. Gendered pronouns occur quite frequently – mostly in the context, less so in the answers, rarely in the questions. Neopronouns are not featured at all in any form (the one *e* is likely to be a typo).

Nevertheless, SIQa exhibits some gender-inclusive tendencies. The gender-neutral singular reflexive *themself* occurs 74 times across the whole data set, indicating that more usage of gender-neutral *they* is likely to be featured more prominently. Besides, also the choice of included names appears to be fairly inclusive after a cursory qualitative inspection, because many of the named agents in SIQa feature gender-neutral names like *Alex* or *Kai*. Yet, it is still rather conservative and does not feature any neopronouns.

4 Benchmark creation

4.1 Neopronouns and (co)reference

Pronouns usually either substitute for a noun (phrase) or are used to signal reference to something that can be inferred from the situational context (Quirk et al., 1985). As such, they are ubiquitous in everyday language, but generally do not attract new lexemes because they constitute a closed word class. Novel items are only slowly introduced via grammaticalization (Heine and Song, 2011). Neopronouns, then, present a unique case; some of them developed organically within specific social groups to promote gender inclusivity, others were deliberately created for that purpose (e.g., *ey* in 1975, *thon* in the late 19th century, see McGaughey, 2020). While neopronouns are gaining traction in some communities, they remain less widely adopted, with gender-neutral *they* being the notable exception.

For SlayQA, we specifically filter out examples that do not include at least two coreference chains with named entities. This filtering is crucial because without multiple entities in the text, the replacement of pronouns with neopronouns does not significantly alter the amount of generalization measured by the task. When only one entity is present, changed pronouns do not pose an insurmountable challenge to a model's understanding of the situation – there are no options for interpreting the neopronoun in/correctly. However, when multiple named entities are involved, the task becomes much more demanding, as the model must accurately track and resolve these coreferences across texts. This ensures that SlayQA actually tests the ability to handle neopronouns *in use*.

4.2 Creating the distribution shift

To create the envisioned distribution shift, we first parsed all examples in the SIQa training and development data sets as a combined `context + question + answers` string with `spacy` (Honnibal et al., 2020) and performed coreference resolution with `coreferee` (Hudson, 2023). For the following data modification step, we included all sentences that feature at least two coreference chains which resolve to proper nouns, i.e. names in the case of the SIQa data set. Sentences without any pronouns or with coreference that resolves to a singular entity were therefore discarded. From the original 35.364 entries in the training data, 1.985 examples were left after this procedure.

In a second step, we then iterated over all leftover examples and filtered out those that did not contain any male or female gendered pronoun. After this step, 1.388 examples were left. For each context-question-answers entry in the filtered data, we then replaced all forms of established male pronouns (forms of *he*) and established female pronouns (forms of *she*) with one randomly chosen set of corresponding neopronoun forms. We decided not to alter forms of *they* as they are a) already used in a gender-neutral fashion in several examples in SIQa, and b) proved to be hard to correctly parse into singular or plural forms, where replacement of the plural form with a singular neopronoun might create illogical examples.

Finally, we noticed that a minority of data points in SIQa feature incorrect or mixed pronoun use. In the following example, *Kai* is first referred to with

<table>
<tr><th colspan="12">Motivation</th></tr>
<tr><td colspan="3"><i>Practical</i></td><td colspan="3"><i>Cognitive</i>
□</td><td colspan="3"><i>Intrinsic</i></td><td colspan="3"><i>Fairness</i>
□</td></tr>
<tr><th colspan="12">Generalisation type</th></tr>
<tr><td colspan="2"><i>Compositional</i>
□</td><td colspan="2"><i>Structural</i></td><td colspan="2"><i>Cross Task</i></td><td colspan="2"><i>Cross Language</i></td><td colspan="2"><i>Cross Domain</i></td><td colspan="2"><i>Robustness</i></td></tr>
<tr><th colspan="12">Shift type</th></tr>
<tr><td colspan="3"><i>Covariate</i>
□</td><td colspan="3"><i>Label</i></td><td colspan="3"><i>Full</i></td><td colspan="3"><i>Assumed</i></td></tr>
<tr><th colspan="12">Shift source</th></tr>
<tr><td colspan="3"><i>Naturally occuring</i></td><td colspan="3"><i>Partitioned natural</i></td><td colspan="3"><i>Generated shift</i>
□</td><td colspan="3"><i>Fully generated</i></td></tr>
<tr><th colspan="12">Shift locus</th></tr>
<tr><td colspan="3"><i>Train–test</i></td><td colspan="3"><i>Finetune train–test</i></td><td colspan="3"><i>Pretrain–train</i></td><td colspan="3"><i>Pretrain–test</i>
□</td></tr>
</table>

Table 3: GenBench Evaluation Card for SlayQA

male pronouns (*himself*, *him*) in the context, but then with the gender-neutral singular *they* in the first answer:

- `context: Tracy saw Kai standing there by himself and decided to go talk to him.`
- `question: How would Kai feel as a result?`
  - `answerA: upset they had to deal with someone`
  - `answerB: they want to be left alone`
  - `answerC: happy to not be lonely anymore`

We acknowledge this haphazard noise in the original data but due to the rarity of its occurrence, we do not further attempt to clean the data from it.

5 SlayQA in the generalisation taxonomy

Table 3 shows where SlayQA is located in the generalisation hierarchy by Hupkes et al. (2023).

Motivation SlayQA is both cognitively and fairness-motivated. Humans are generally able to use neopronouns correctly and productively. Consequently, if language models are indeed good models of human language (usage), they should not struggle with social reasoning that includes neopronouns. Additionally, the inclusion and correct processing of neopronouns also relates to fairness of language technologies – they should be applicable to all potential users, even in the light of a changing linguistic and societal landscape (cf. also related work in Section 2).

Generalisation type SlayQA assesses whether language models can interpret novel, highly infrequent pronoun forms in social reasoning contexts. This is a test for compositional generalisation, as neopronouns are systematic, productively used, substitutive with regard to the referents they replace, and localist in the sense of only depending on context, question and answer sentences. As such, our benchmark fulfils the criteria essential for compositional generalisation, as laid out by Hupkes et al. (2020).

Shift type Our benchmark constitutes a covariate shift. We assume that social reasoning of the kind that SIQa tests is somehow implicitly, if not explicitly, included in the pre-training data. By changing the pronouns in the complete data sets (context, questions and answers), we do not alter the nature of the task or the correctness patterns of the answers. While the test distribution now differs more strongly from the training distribution: $p(x_{tst}) \neq p(x_{tr})$, the conditional probabilities still stay the same: $p(y_{tst}|x_{tst}) = p(y_{tr}|x_{tr})$.

Shift source Our benchmark includes a generated shift. As the original SIQa data set is crowdsourced, it is reasonable to assume that it still follows a somewhat *representative*, if not completely *authentic* (in the sense of Stefanowitsch, 2020) linguistic distribution, comparable to common pre-training corpora without synthetic data. This representative distribution is explicitly altered for SlayQA by including a much higher proportion of neopronouns than classical corpora.

Shift locus The data shift is localized between pre-training and testing. We explicitly do not fine-tune the models on social reasoning with neopronouns, as we are interested in the compositional abilities of LLMs *as is* and do not want to skew them with additional training on data that reflects the fairness generalization we aim to assess.

6 Evaluation

6.1 Methodology

Models We evaluate five different, autoregressive models: OLMo-1B (Groeneveld et al., 2024) as a representative model for the Dolma pre-training corpus (Soldaini et al., 2024), three RedPajama-INCITE-7B models (base, chat-tuned and instruction-tuned) for the RedPajama-Data-1T data (Together Computer, 2023) and a quantized version of the instruction-tuned Llama-3.1 8B (Team, 2024) for C4 (Raffel et al., 2020). Because the original Llama-1 was provably trained on C4 (among other data sets, see Touvron et al., 2023a), we assume that this data set is still fully present in the training data of Llama-3. However, it is not clear whether this is actually the case, since the most recent Llama paper (Team, 2024) does not reveal any information about the concrete make-up of the pre-training data.

Due to the exorbitant resource demands of so-called small state-of-the-art models like OlMo-7B or Llama-3.1-8B, their evaluation on this proposed benchmark was, unfortunately, beyond the capabilities of our available GPU resources. Therefore, we opted to only evaluate smaller (OLMo-1B) or quantized models. For the Llama-3.1 model, we had to resort to a version working with lower number precision (quantized from FP16 down to INT4 with `AutoAWQ`, based on Lin et al., 2024). Unfortunately, this specific configuration is only available for the instruction-tuned model, so we do not provide scores for the base model.

Data We evaluate our models on three data sets: our distribution-shifted benchmark, the original 1.388 unaltered data points with at least two coreference chains, and a random selection of 1.388 examples sampled from Social IQa that were not restricted with regard to coreference. For reproducibility reasons, we host SlayQA[1], the randomly sampled Social IQa set[2] and the unaltered data points, NoSlayQA[3], on the Hugging Face hub.

Scoring To assess the preference of the individual models, we use the Hugging Face `transformers` library (Wolf et al., 2020) and its evaluation metrics. In line with Brown et al. (2020), we measured the language models' preference for a specific answer by calculating its probability conditioned on the context and question. To do so, we chose a perplexity-based (Jelinek et al., 1977) approach.[4] We calculated the perplexities of concatenated context + question + answer strings for all three choices in each example and then selected the answer with the lowest perplexity as the one preferred by the model. As such, we perform zero-shot evaluation on models not explicitly fine-tuned for this task. Performance is measured as accuracy against the gold standard labels in the data.

6.2 Results

The results of the zero-shot evaluation are displayed in Table 4. Across all models and evaluation data sets, the results lie between 7% and 13% above the baseline. This indicates that all models have acquired some generalization capabilities in the social reasoning domain, at least as instantiated by the SIQa/SlayQA question patterns. The differences between the models and between the data sets for the various models are comparatively small, but still exhibit somewhat systematic patterns.

From a model-centric viewpoint, the Llama-3.1-8B model in particular outperforms the other models, achieving the highest scores on all three data sets – 46.97% on the SIQa subset, 46.4% on NoSlayQA, and 44.16% on SlayQA. This quantized model consistently outpaces the RedPajama series and the OLMo-1B model by three to four percentage points. The RedPajama models demonstrate slightly varying performance, with the Base variant surpassing the others on the SIQa subset and the instruction-tuned version achieving the worst performance. The scores for OLMo-1B are comparable to the best RedPajama scores.

[1] `https://huggingface.co/datasets/bbunzeck/slayqa`

[2] `https://huggingface.co/datasets/bbunzeck/minisiqa`

[3] `https://huggingface.co/datasets/bbunzeck/noslayqa`

[4] Original experiments with the `outlines` library (Willard and Louf, 2023) and constrained generation showed similar tendencies, but generally resulted in lower accuracy scores.

Model	SIQa subset	NoSlayQA	SlayQA
Random baseline	33.33%	33.33%	33.33%
allenai/OLMo-1B	43.88%	42.87%	42.21%
togethercomputer/RedPajama-INCITE-Base-3B-v1	42.87%	43.3%	40.99%
togethercomputer/RedPajama-INCITE-Instruct-3B-v1	40.71%	41.35%	40.78%
togethercomputer/RedPajama-INCITE-Chat-3B-v1	41.5%	43.95%	41.93%
hugging-quants/Meta-Llama-3.1-8B-Instruct-AWQ-INT4	46.97%	46.4%	44.16%

Table 4: Results for different models

When comparing the data sets, it is striking that scores for SlayQA are consistently lower than the scores the unaltered NoSlayQA. Despite the differences being fairly marginal, this pattern is stable across all five models. Interestingly, performance on the (presumably easier) SIQa subset, which was not filtered for two coreference chains, is not always higher than the performance on (No)SlayQA. While this is the case for OLMo-1b and Llama-3.1-8B, the RedPajama models always perform better on NoSlayQA than on the SIQa subset.

7 Discussion and conclusion

We created SlayQA as a more inclusive benchmark for evaluating question answering and social reasoning in LMs. A key motivation was to test how well these models can generalize in this domain, particularly under the significant distribution shift between pre-training and test data that we created by replacing gendered pronouns with neopronouns that occur much less frequently in the investigated pre-training corpora.

The results indicate that we have succeeded in our objectives. The scores on SlayQA are consistently lower than those on the parallel NoSlayQA data set, which suggests that the models struggle more with the challenges SlayQA presents. Interestingly, however, the scores on SlayQA are not always below our random selection SIQa data. This finding is intriguing because we expected questions requiring the tracking of two coreference chains to be more challenging than those without such demands. It is quite possible that questions without two coreference chains introduce different, perhaps equally complex, challenges. Moreover it should also be noted that the relatively small differences between models could be due to training noise – for an even more comprehensive evaluation of neopronouns' influence, several comparable models that differ in their random initializations would be needed. As these are not readily available and costly to train, we have to leave this direction to future work.

From a model-centric standpoint, the largest model (Llama-3.1-8B) consistently outperforms the smaller models, even though it was drastically quantized to much lower number precision. Among the smaller models, the performance differences are minimal, with no substantial gap between the 1B OLMo and 2B RedPajama models. Additionally, there are no significant differences between the base RedPajama model and those fine-tuned for instruction-following or conversational tasks, which is surprising given the assumption that fine-tuning should improve performance on question answering tasks compared to vanilla models.

Although we were not able to evaluate larger models, our accuracies do not deviate drastically from comparable zero-shot evaluations for much larger models. In the the Llama-1 paper, Touvron et al. (2023a) report scores between 48.5% for Llama-1-7B and 52.3% for Llama-1-65B. For Llama-2 (Touvron et al., 2023b), scores align as well (48.3% for the 7B model, 50.7% for the 70B model). Even the largest Llama-3 model with 405B parameters only achieves 53.7% on SIQa, as reported in Team (2024). Judging from these meagre scaling effects, we assume that evaluations of larger models on SlayQA should not deviate drastically.

While we decided to employ a zero-shot evaluation approach for comparability, it would also be interesting to see how models fare in multi-shot reasoning or fine-tuning contexts. The AllenAI leader board for SIQa[5] reports the best fine-tuned model with a score of 84.31%. Furthermore, prompting has started to replace more technical evaluation ap-

[5] https://leaderboard.allenai.org/socialiqa/submissions/public

proaches (although it remains debated, see Hu and Levy, 2023) – as such it would be also interesting to see how commercial and open models work in SIQa in prompting settings with chain-of-thought or different reasoning approaches.

Future research similar to SlayQA should definitely aim to include even more novel and linguistically interesting forms, e.g. the aforementioned nounself, emojiself, numberself or nameself pronouns (Lauscher et al., 2022). Their unique structure and usage should be even rarer than neopronouns and could pose even more veritable challenges to the generalization capabilities of modern LMs. Additionally, the SlayQA paradigm could be expanded to other benchmarks that test different capabilities. For example, it would be interesting whether performance on the grammatical benchmark BLiMP (Warstadt et al., 2020) deteriorates with the inclusion of neopronouns. Finally, the influence of different language modeling choices, e.g. tokenization, deserves further scrutiny. As our tested models did not drastically differ in subword tokenization for the tested (neo)pronouns, we cannot draw definite conclusions. Evaluation on a wider range of models could illuminate this further.

Limitations

As this study is the first of its kind, it is still limited in various ways. As previously mentioned, evaluations of extremely large LMs were impossible due to limited resources. Yet, the comparison of our results with those of contemporary model reports showed similar scores, so we assume that this limitation did not impact the current study in a major way. Another limiting factor lies in the focus on neopronouns. As mentioned in the previous paragraph, further ways of gender-inclusive pronoun usage exist. Each one of them is deserving of recognition in inclusive NLP research, but for the sake of brevity, we focused on neopronouns only. As a final limitation, we question the quality of SIQa as a benchmark of social reasoning. We include one example in section 4.2, but our manual inspection of SIQa yielded many more such examples with confusing or borderline nonsensical "correct" answers. While we chose it as the base of SlayQA due to its widespread use in evaluation of SOTA models, its quality for the kind of reasoning it aim to evaluate is fairly questionable.

Acknowledgements

We thank Clara Lachenmaier and Judith Sieker for their helpful comments and remarks on earlier drafts of this paper.

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC-1646, project number 512393437, project A02.

References

Y Gavriel Ansara and Peter Hegarty. 2013. Misgendering in English language contexts: Applying noncisgenderist methods to feminist research. *International Journal of Multiple Research Approaches*, 7(2):160–177.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. 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. Association for Computational Linguistics.

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*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's In My Big Data? *Preprint*, arXiv:2310.20707.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the Science of Language Models. *Preprint*, arXiv:2402.00838.

Bernd Heine and Kyung-An Song. 2011. On the grammaticalization of personal pronouns. *Journal of Linguistics*, 47(3):587–630.

Laura Hekanaho. 2021. Generic and Nonbinary Pronouns: Usage, acceptability and attitudes. *Neuphilologische Mitteilungen*, 121(2):498–509.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of Large Language Models in Understanding Pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5352–5367, Toronto, Canada. Association for Computational Linguistics.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Richard Paul Hudson. 2023. Coreferee.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality Decomposed: How do Neural Networks Generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. A taxonomy and review of generalization research in NLP. *Nature Machine Intelligence*, 5(10):1161–1174.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anne Lauscher, Debora Nozza, Ehm Miltersen, Archie Crowley, and Dirk Hovy. 2023. What about "em"? How Commercial Machine Translation Fails to Handle (Neo-)Pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–392, Toronto, Canada. Association for Computational Linguistics.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100.

Sebastian McGaughey. 2020. Understanding Neopronouns. *The Gay & Lesbian Review Worldwide*, 27(2).

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1246–1266, Chicago IL USA. ACM.

Anaelia Ovalle, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard Zemel, Aram Galstyan, Yuval Pinter, and Rahul Gupta. 2024. Tokenization matters: Navigating data-scarce tokenization for gender inclusive language technologies. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1739–1756, Mexico City, Mexico. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *ACM Computing Surveys*, 55(10):1–45.

Ell Rose, Max Winig, Jasper Nash, Kyra Roepke, and Kirby Conrod. 2023. Variation in acceptability of neologistic English pronouns. *Proceedings of the Linguistic Society of America*, 8(1):5526.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense Reasoning about Social Interactions. 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 4462–4472, Hong Kong, China. Association for Computational Linguistics.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *Preprint*, arXiv:2402.00159.

Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. *Preprint*, arXiv:2112.14168.

Anatol Stefanowitsch. 2020. *Corpus Linguistics: A Guide to the Methodology*. Language Science Press, Berlin.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Llama Team. 2024. The Llama 3 Herd of Models. Technical report.

Together Computer. 2023. RedPajama: An open dataset for training large language models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. *Preprint*, arXiv:2307.09288.

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.

Brandon T. Willard and Rémi Louf. 2023. Efficient Guided Generation for Large Language Models. *Preprint*, arXiv:2307.09702.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick Von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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.

A Pronoun frequencies in pre-training corpora

The following tables contain the absolute token frequencies of all non-subject forms of the established gendered pronouns *he* and *she*, the gender-neutral *they* (which is more commonly used in the plural form, except the explicitly singular *themself*), and eight neopronouns from Hossain et al. (2023). Frequencies are reported for C4 (Raffel et al., 2020) and Dolma (Soldaini et al., 2024), and were calculated with WIMBD (Elazar et al., 2024).

	C4	Dolma
him	76.642.827	466.261.554
her	120.502.480	610.920.325
them	206.400.522	1.224.786.435
thon	872.654	992.499
em	14.687.464	25.071.924
aer	607.125	638.705
co	83.935.707	199.206.147
vir	456.939	645.878
xem	285.577	357.204
em	14.687.464	25.071.924
zir	22.433	40.578

Table 5: Frequencies for established object pronouns and object neopronouns in C4 and Dolma

	C4	Dolma
his	154.746.745	932.171.598
her	120.502.480	610.920.325
their	300.195.337	1.677.918.677
thons	54.734	90.213
es	17.287.828	20.223.489
aer	607.125	638.705
cos	2.040.163	5.310.600
vis	2.335.366	4.775.286
xyr	3.579	10.039
eir	201.341	375.303
zir	22.433	40.578

Table 6: Frequencies for established possessive determiners and neo-determiners in C4 and Dolma

	C4	Dolma
his	154.746.745	932.171.598
hers	2.652.526	10.223.659
theirs	2.429.259	12.494.222
thons	54.734	90.213
ems	2.938.663	4.043.142
aers	20.147	24.815
cos	2.040.163	5.310.600
virs	4.374	11.125
xyrs	1.912	1.977
eirs	20.911	24.996
zirs	681	2.317

Table 7: Frequencies for established possessive pronouns and possessive neopronouns in C4 and Dolma

	C4	Dolma
himself	22.378.650	134.674.595
herself	11.941.936	63.594.961
themself	158.315	1.289.078
thonself	36	248
emself	1.017	2.341
aerself	28	161
coself	51	193
virself	0	49
xemself	155	626
emself	1.017	2.341
zirself	387	1.695

Table 8: Frequencies for established reflexive pronouns and reflexive neopronouns in C4 and Dolma

B Token numbers for (neo)pronouns

	OLMo					RedPajama-INCITE					Llama-3.1-8B				
	Subj.	Obj.	Det.	Poss.	Reflex.	Subj.	Obj.	Det.	Poss.	Reflex.	Subj.	Obj.	Det.	Poss.	Reflex.
he	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
she	1	1	1	1	2	1	1	1	1	2	1	1	1	2	3
they	1	1	1	2	2	1	1	1	2	2	1	1	1	2	2
thon	1	1	2	2	3	1	1	2	2	3	1	1	2	2	2
e	1	1	1	1	2	1	1	1	1	2	1	1	1	1	2
ae	1	2	2	2	3	1	2	2	2	3	1	2	2	2	3
co	1	1	1	1	2	1	1	1	1	2	1	1	1	1	2
vi	1	1	1	2	2	1	1	1	2	2	1	1	1	2	2
xe	1	2	2	2	3	1	2	2	2	3	1	2	2	2	3
ey	1	1	2	2	2	1	1	2	2	2	1	1	2	2	2
ze	1	2	2	2	3	1	2	2	2	3	1	2	2	2	3

Table 9: Number of sub-word tokens that a form of a (neo)pronoun is split into by a specific model