What Are They Filtering Out? An Experimental Benchmark of Filtering Strategies for Harm Reduction in Pretraining Datasets

Marco Antonio Stranisci[1,2], **Christian Hardmeier** [3]
[1]Università degli Studi di Torino
[2]aequa-tech
[3] IT University of Copenaghen
marcoantonio.stranisci@unito.it

Abstract

Data filtering strategies are a crucial component to develop safe Large Language Models (LLM), since they support the removal of harmful contents from pretraining datasets. There is a lack of research on the actual impact of these strategies on vulnerable groups to discrimination, though, and their effectiveness has not been yet systematically addressed. In this paper we present a benchmark study of data filtering strategies for harm reduction aimed at providing a systematic evaluation on these approaches. We provide an overview 55 technical reports of English LMs and LLMs to identify the existing filtering strategies in literature and implement an experimental setting to test their impact against vulnerable groups. Our results show that the positive impact that strategies have in reducing harmful contents from documents has the side effect of increasing the underrepresentation of vulnerable groups to discrimination in datasets.

Code — https://github.com/marcostranisci/benchmarking-strategies

Datasets — https://huggingface.co/datasets/marco-stranisci/people-dataset

1 Introduction

The harmfulness of Large Language Models (LLM) is an open issue that gathers the attention of different sectors of our society. International bodies regulated the development of these technologies (Edwards 2021); Natural Language Processing (NLP) scholars are introducing a series of approaches (Touvron et al. 2023; Dutta et al. 2024) to assess and mitigate their impact against vulnerable groups to discrimination.

Even if the development cycle of a LLM encompasses several steps, in recent years most of the research focuses on the post-training stage, for which several benchmarks (Gehman et al. 2020; Tedeschi et al. 2024) have been created. Theoretical research on effective strategies to filter out harmful contents from pretraining datasets is an understudied topic, though. If compared to the amount of LLMs released in recent years, only a limited number of approaches to filtering strategies has been proposed (Raffel et al. 2020; Brown et al. 2020), and many o them have been implemented without considering the complex nature of bias, producing unwanted negative effects against several categories of people (Dodge et al. 2021; Xu et al. 2021). Few critical studies on filtering strategies (Luccioni and Viviano 2021; Longpre et al. 2024) have been performed so far, but none of them systematically addresses the topic.

The aim of our research is to propose the first systematic analysis of data filtering strategies for harm reduction in pretraining dataset. Specifically, we formulate two research questions.

RQ 1: Which filtering strategies are implemented to remove harmful contents from pretraining datasets? We surveyed 55 technical reports describing English LMs and LLMs to collect information about the characteristics of the existing data filtering strategies and their documentation. Through the survey we have been able to identify eight different categories of filtering strategies for harm reduction that have been proposed in literature. The survey also shows a disengagement trend in current LLM technical reports, from which emerges a general lack of awareness on their role in increasing the underrepresentation of vulnerable groups to discrimination in datasets.

RQ 2: Which categories of people are most affected by filtering strategies? We performed a benchmark analysis on seven data filtering strategies to evaluate if and to which extent they increase the underrepresentation of vulnerable groups to discrimination in pretraining datasets. To perform this analysis, we designed a pipeline for identifying the mentions of named entities categorized by their gender and origin. Results of our analysis show that women are the most impacted by filtering strategies and that strategies significantly differ in the contents they filter out. Choosing a specific strategies means focusing on specific sources of harm overlooking others.

The contribution of our work is threefold: *i.* we provide the first systematic survey of data filtering strategies for harm reduction; *ii.* we implement a pipeline and a set of resources to benchmark filtering strategies; *iii.* we test our pipeline on a pool of data filtering strategies, empirically demonstrating the negative impact of these approaches in the underrepresentation of vulnerable groups to discrimination.

2 Related Work

Research on data filtering strategies for harm reduction has been propelled by the research on biases in LLMs (Bender et al. 2021) and in pretraining datasets (Jo and Gebru 2020). However, critical studies on strategies focused on specific datasets rather than the impact of strategies themselves. Luccioni and Viviano (2021) analyzed the presence of Hate Speech in the Common Crawl dataset. Dodge et al. (2021) documented the C4 corpus (Raffel et al. 2020), showing that its filtering process correlates with a reduction of terms defining vulnerable groups to discrimination in datasets. (Longpre et al. 2024) tested the correlation between removing toxic contents from pretraining datasets and LLMs performance in toxicity classification, showing that filtering has a negative impact on this task. The survey of Albalak et al. (2024) on data selection for LLMs provide a marginal and non-systematic taxonomy of strategies. Finally, the work of (Lucy et al. 2024) on language filtering strategies provide insightful results on the cultural biases embedded in language filters that can determine the exclusion of categories of people who speak non-standard English varieties.

Our work fills the existing gap providing a first systematic analysis of the impact of filtering strategies for harm reduction in datasets.

3 An Overview of Existing Data Filtering Strategies

In this section we present our systematic overview of data filtering strategies for harm reduction in pretraining data as they are presented in 55 LMs technical reports. We first introduce our methodology for the selection of relevant documents and their analysis. Then, we present a taxonomy of eight existing data filtering strategies that emerges from the analysis. Finally, we identify some trends emerging from a general overview of reports in a diachronic perspective.

3.1 Overview Methodology

To ensure a representative pool of technical reports describing LLMs, we seeded them from six leaderboards that have been chosen to include in our study different generations of LMs as well as different tasks. To ensure that small LMs were included in our survey we gathered all the technical reports of systems ranked higher than the baseline of SuperGLUE (Wang et al. 2019) and the first 50 highest-ranked in SQuAD (Rajpurkar et al. 2016). LLMs technical reports were collected from MMLU-pro (Wang et al. 2024), from which we gathered all systems that have an aggregated score that is equal or higher than 0.5, and the first 50 high-ranked from Chatbot Arena (Chiang et al. 2024). Finally, we included all models that have been benchmarked in two existing leaderborards focused on LMs safety: ALERT (Tedeschi et al. 2024) and Secure Learning Lab's LLM safety leaderboard [1]. In this first step we collected 47 technical reports.

For each report we checked three types of content: the description of the pretraining dataset, the presentation of data filtering strategies, and the treatment of harmful or biased contents in datasets. In the context of this research, we adopt the term harmful for all the phenomena that might have a negative impact against vulnerable groups to discrimination. In this sense we refer to survey of (Blodgett et al. 2020), who distinguish between *representational* harms, which consist in all the negative representation of groups, and *allocative* harms, which includes all the forms of underrepresentation.

For each paper we search the following keywords: *data, filter, quality, toxic, bias, hate, stereotype*. This stage led to the retrieval of 13 additional papers referenced in the sections that cover the description of filtering strategies. Among them, 4 describe the creation of a pretraining dataset and 1 a methodology for data sampling. At the end of this process we obtained a pool of 55 papers for our survey with information about the type of data filtering strategies that have been implemented to remove harmful contents and their availability.

3.2 A Taxonomy of Data Filtering Strategies

We defined a taxonomy to categorize filtering strategies in Table 1. Given the presence of different LMs belonging to the same family (e.g., LLama, ERNIE), we only listed in the table the latest release of this set of model, unless there are significant changes in data filtering strategies. A change in data filtering is the reason why GPT-4 (Achiam et al. 2023) is treated separately from GPT-3 (Brown et al. 2020) while the previous two instantiations of the model are not mentioned. Whenever the technical report of an LM referred to a dataset created in the context of developing the LLM, such as Olmo (Groeneveld et al. 2024) with the Dolma corpus (Soldaini et al. 2024), we jointly mentioned them in the table.

Authoritative sources. This strategy is based on the selection of documents only from authoritative resources validated from the research community. Even if this strategy relies on selection rather than filtering, we mentioned it since it represents a standard approach before the wave of research on the ethical issues related to language modeling (Bender et al. 2021). An example of this approach is in BERT (Kenton and Toutanova 2019) that has been trained on a snapshot of Wikipedia and on the Book Corpus (Zhu et al. 2015) without implementing any type of quality checks and filters.

Document seeding. This strategy is based on heuristics for the selection of specific documents from the web. One of the most famous examples of this approach is the one adopted for the creation of the OpenWebText corpus (Gokaslan et al. 2019), which is a collection of all the outbound links from Reddit that have been upvoted at least 3 times.

Quality-based. This strategy filters out low-quality documents by comparing them with high-quality documents selected from authoritative sources. For instance, Brown et al. (2020) created a corpus of documents sampled from Wikipedia, OpenWebtext, and the Book Corpus to train a classifier for the removal of documents that are classified as too dissimilar from it.

[1] https://huggingface.co/spaces/AI-Secure/llm-trustworthy-leaderboard

strategy	models
authoritative source	BERT (Kenton and Toutanova 2019), RoBERTa (Liu 2019), XLNet (Yang 2019), *DeBERTa* (He et al. 2021), ERNIE 3.0 (Sun et al. 2021), **LinkBERT** (Yasunaga, Leskovec, and Liang 2022), *Vega v2* (Zhong et al. 2022)
document seeding	*GPT-3* (Brown et al. 2020), **DSIR** (Xie et al. 2023), Mamba (Gu and Dao 2023), MPT-7B (Research 2023), MAmmoTH2 (Mickus et al. 2024)
quality-based	*GPT-3*(Brown et al. 2020), Gopher (Rae et al. 2021), GLaM (Du et al. 2022), Mamba (Gu and Dao 2023), MAmmoTH2 (Mickus et al. 2024)
toxicity classifier	Qwen (Bai et al. 2023), Gemma (Team et al. 2024a), **OLMO-Dolma** (Soldaini et al. 2024), Yi-Lighting (Wake et al. 2024)
rule-based	**T5-C4** (Raffel et al. 2020), *FALCON* (Penedo et al. 2023), Gemma (Team et al. 2024a)
url blacklists	*FALCON* (Penedo et al. 2023),
human-in-the-loop	LaMDA (Thoppilan et al. 2022), **Bloom-ROOTS** (Laurençon et al. 2022), **Phi-3.5** (Abdin et al. 2024)
safety policy	Gemini (Gemini et al. 2023), EXAONE 3.5 (Research et al. 2024), **INCITE-RedPajama** (Weber et al. 2024), LLama-3 (Dubey et al. 2024)
not mentioned	Alpaca (Taori et al. 2023), Claude (Antropic 2023), GPT-4 (Achiam et al. 2023), Zephyr (Tunstall et al. 2023), DeepSeek-V3 (DeepSeek-AI et al. 2024), Grok-2 (team 2024), Jamda (Lieber et al. 2024), Mixtral (Jiang et al. 2024), Nova (Intelligence 2024), Reka (Team et al. 2024b)

Table 1: A taxonomy of data filtering strategies described in technical reports describing LLMs. Model names are highlighted with different text styles depending on the availability of filtering strategies and/or dataset: **bold** if they are fully available, *italics* if they are partially available, underlined if they are not available.

Toxicity classifier. This strategy provides the training of a classifier to filter out all the potentially harmful contents from datasets. The first example of such an approach in our survey is from Gao et al. (2020) who used profanity-check[2] to remove toxic contents from The Pile dataset. Other commonly adopted classifiers are Perspective APIs[3], and FastText models trained on corpora for Hate Speech (HS) detection (Soldaini et al. 2024).

Rule-based. This strategy exploits lexicons or heuristics to remove unwanted contents. The most studied strategy of this type adopts the Shutterstack Lexicon[4] (Raffel et al. 2020) to flag as toxic any sentence containing one of the word in that list.

Url blacklist. This strategy removes contents from blacklisted websites. The strategy has been adopted for the creation of the RefinedWeb Dataset (Su et al. 2024), which relied on the semi-automatic creation of a blacklist of urls.

Human-in-the-loop. This strategy provides the involvement of human evaluators during the creation of datasets. Laurençon et al. (2022) organized hackatons with NLP communities to create a whitelist of domains to be used to crawl data for pretraining. Thoppilan et al. (2022) collected human prompts from crowd-workers aimed at collecting a corpus of Question Answering pairs annotated for harm detection.

Safety policy. These strategy relies on the self-assessment of the research team who develops the LM. For instance, the team that developed Gemini declare that they "perform safety filtering to remove harmful content based on our policies" (Gemini et al. 2023). Dubey et al. (2024) detect toxic contents without removing them and use such an information to implement their safety policy in a further step of their language modeling pipeline.

3.3 Trends and Limitations in Data Filtering

In this section we discuss some general trends and limitations that emerge from our overview.

A first consideration is about the **the lack of replicability** of filtering strategies presented in technical reports. This trend is shown in Table 1 where models are underlined if they did not disclose their data filtering pipeline or the snapshot of datasets used for training, written in *italics* if they partially did it, in **bold** if they fully disclose filtering strategies and data. As can be observed, most of the existing LMs are not released with the actual scripts for the replication of implemented filtering methods. Specifically, we recognize a first generation of LMs (e.g.: BERT, RoBERTa, etc.) that have been trained without considering the issue of harmful contents in pretraining datasets. The pivotal work of Bender et al. (2021) led to a second generation of models that

[2]https://pypi.org/project/profanity-check/

[3]https://perspectiveapi.com/

[4]https://bit.ly/3QjwMvz

implemented specific strategies and open-sourced their approaches and results for further investigation. This is the case of the C4 corpus (Raffel et al. 2020), which has been released both in a filtered and unfiltered version. Besides notable exceptions (Soldaini et al. 2024; Weber et al. 2024), the current tendency is to not disclose scripts and results of the data filtering strategies implemented for pretraining.

A similar trend towards the reduced openness of filtering strategies is observable about the **documentation debt** (Bandy and Vincent 2021) of implemented filtering strategies. If the introduction of documentation templates like Datasheets (Gebru et al. 2021) and Model Cards (Mitchell et al. 2019) led to a first wave of fully documented self-assessment reports (Chowdhery et al. 2023), the more recent technical reports formally rely on these documentation templates but substantially provide very generic descriptions about their data filtering approach. For instance, the description provided in the Gemma technical report is limited to the following paragraph: "We filter the pre-training dataset to reduce the risk of unwanted or unsafe utterances [...]. This includes both heuristics and model-based classifiers to remove harmful or low-quality content" (Team et al. 2024a).

A final consideration is about the **lack of strategies that are focused on reducing the underrepresentation of groups vulnerable to discrimination** in pretraining datasets. Although it has been shown that filtering strategies for harm reduction have a negative impact on minorities (Dodge et al. 2021), few attempts to mitigate this issue have been made so far. A notable exception is the work of Soldaini et al. (2024), which provides an assessment of their toxicity classifier over different types of English dialects. However, research works that systematically study the correlation between filtering strategies and increase of minorities underrepresentation have not yet been provided. In Section 4 we present a first experiment focused on this issue.

4 Measuring the Impact of Filtering Strategies Against Vulnerable Groups

In this section we benchmark seven filtering strategies on their impact in reducing or increasing the representativeness of vulnerable groups in pretraining datasets. Our experimental setup adopts an intersectional approach (Crenshaw 2013) as it considers four groups derived from the intersection of people's gender and origin: Western men, Western women, Post-colonial men, Post-colonial women. For this analysis we measure the number of named entities for each demographic group that are removed by the implementation of different filtering strategies, using samples of documents gathered from Common Crawl as a benchmark. In Section 4.1 we describe our experimental setup, in Section 4.2 we discuss the results of our the analysis

4.1 Experimental Setting

Knowledge base creation. The first step of our experimental setting was the creation of a knowledge base that enables the categorization of named entities on the basis of their gender and origin. In order to do so we developed the **People Dataset**, a corpus of 10.8 million of entities of the type person with information about their country of birth, citizenship, gender, ethnic minority, and occupation. The dataset is derived from Wikidata (Erxleben et al. 2014), a collaborative Knowledge Graph (KG) maintained by the Wikimedia ecosystem, and processed on the basis of previous literature on the topic (Stranisci et al. 2023b): we inferred people's countries of birth from Wikidata property 'place of birth' (P19) properties, mapped all the properties of the type 'ethnic group' (P172) with a curated list of Post-colonial minorities in Western countries (e.g., African-Americans). Finally, we integrated the knowledge base with additional information from CaLiGraph (Heist and Paulheim 2019), a KG derived from Wikipedia categories.

Sampling from Common Crawl. We gathered 5 samples of documents from the Common Crawl (CC) snapshot released in August 2024[5]. Each sample is composed of 20 Web ARChive (WARC) files that have been processed according to the following criteria: *i.* we kept only documents classified as written in English, implementing the langdetect Python library[6]; *ii.* we removed all the documents with an average number of sentences below 5 and an average number of words *per* sentence below 5. The average number of documents in each sample is $102,820$ (std: 443).

Entity Linking. The next step of our approach was the linking of named entities with the People Dataset. In order to adopt a computationally-efficient Entity Linking (EL) approach to implement on a large set of data, we designed an EL pipeline that combines the adoption of neural methods for the detection of named entities and existing heuristics to link them to the dataset (Stranisci et al. 2023a). The pipeline was organized in four steps.

1. We used the largest SpaCy model[7] to detect all the named entities of the type PERSON in CC samples.
2. We queried Wikipedia APIs[8] with all the retrieved named entities to obtain titles of Wikipedia pages and filtered out all the named entities that were too dissimilar from Wikipedia titles, through the adoption of a set of heuristics already validated by Manghi (2023).
3. We retrieved all the Wikidata IDs corresponding to the Wikipedia titles.
4. We kept only entities with a Wikidata ID that is present in our People Dataset.

As a result (Table 2), we identified an average of $128,025$ entities from each sample: $85,276$ Western men, $15,272$ Post-colonial men, $22,185$ Western women, $5,292$ Post-colonial women. In order to assess whether there is high variation between samples we conducted an ANOVA test (St, Wold et al. 1989) on all the possible pairs of entities distributions broken-down by group. In all cases we obtained a $p - value$ of 0.99, showing that there is no evidence of variation between different CC samples.

[5]https://data.commoncrawl.org/crawl-data/CC-MAIN-2024-33/index.html

[6]https://pypi.org/project/langdetect/

[7]https://spacy.io/,en_core_web_sm

[8]https://www.mediawiki.org/wiki/API:Search

	strategy	w.m.	p-c.m	w.w.	p-c. w.
	unfiltered	85276.0	15272.0	22184.6	5292.0
2*rule-based	shutterstack	-2.3%	-1.6%	**-4.2%**	-3%
	hatebase	-0.4%	-0.6%	-0.5%	**-0.9%**
3*classifier-based	perspective	-0.11%	-0.11%	**-0.13%**	-0.11%
	fasttext	-0.3%	-0.3%	**-0.9%**	-0.7%
	profanity-check	-0.21%	-0.22%	**-0.89%**	-0.52%
2*quality-based	quality_wiki	-15.1%	**-17.2%**	-12.7%	-11.2%
	quality_webtext	**-44.6%**	-42.6%	-33.1%	-33.4%

Table 2: The average number of named entities in the 5 samples gathered from Common Crawl and the percentage of sentences removed by applying 7 filtering strategies. Named entities are broken down by groups resulting from the intersection of gender and origin: Western men (w.m.), Post-colonial men (p-c.m.), Western women (w.-w.), and Post-colonial women (p-c.-w.)

Implementation of Filtering Strategies. Having created our baseline, we implemented seven filtering strategies belonging to three different approaches outlined in the taxonomy 1, which can be replicated in an experimental setting: rule-based, toxicity classifier, and document similarity. These strategies were applied to all the sentences or documents that mentioned at least one of the retrieved named entities.

The implementation of rule-based strategies relies on two different lexicons for the identification of toxic language: the Shutterstack lexicon and a version of HateBase that has been refined by (Davidson et al. 2017) from their Hate Speech corpus[9]. We flagged as harmful any sentence that contains at least one term included in each lexicon.

For the comparison of model-based strategies we adopted three toxicity classifiers: Perspective API, which has been used during the training of models like Gemini (Gemini et al. 2023) and Gopher (Rae et al. 2021), profanity-check[10], which has been used to filter the Pile Corpus (Gao et al. 2020), a FastText classifier trained on the Jigsaw dataset (Jain et al. 2022) that has been used to polish Dolma[11] (Soldaini et al. 2024). We considered all sentences classified containing Hate Speech by the FastText classifier as harmful. Since Perspective APIs and profanity-check output a probability distribution, we filtered out all sentences classified as toxic with a probability of 0.8 or more, a threshold in line with previous research (Xu et al. 2021).

For the document similarity strategy we replicated the methodology proposed by (Brown et al. 2020) and (Du et al. 2022). We trained a hash based-linear classifier with a vector-size of $1,000$ on a corpus of high-quality documents to filter out documents that do not reach a given quality threshold. Classifiers were trained on a total of $100,000$ documents: $50,000$ considered as high-quality (positive class), $50,000$ considered as low-quality, which were randomly gathered from CC WARC files that were not used during our sampling step. We chose two sources of high-quality documents: the first based on a random sample of Wikipedia documents, the second on a sample from the OpenWebText corpus (Gokaslan et al. 2019).[12]

4.2 Analysis of Results

Table 2 shows the results of our experiments. The first row of the table reports the average number of entity mentions in samples. In the other rows the impact of a data filtering strategy on named entities is reported in terms of the percentage of removed mentions from the sample. Filtering strategies are grouped by the taxonomy they belong to (Section 3.2) and results are broken-down by sociodemographic group: Western men (w.m.), Western women (w.w.), Post-colonial men (p-c.m.), and Post-colonial women (p-c.w.).

How much is filtered out? From a general overview of results it is possible to observe that strategies have a widely varying impact on our samples. The filtering strategy based on the Hatebase lexicon (Davidson et al. 2017) and all classifier-based approaches have a marginal effect on documents, filtering out less that 1% of sentences. The Shutterstack lexicon has a greater impact, spanning between 1% and 4.2% of removed mentions. Document-similarity approaches lead to a more aggressive filtering that ranges between 11.2% and 44.6% of removed contents.

What is filtered out? The magnitude of filtering strategies' impact is not fully explicative of their behavior without considering which texts they flag for the removal. For each category of filtering strategy we performed an analysis aimed at understanding which pattern they follow for removing contents. To investigate rule-based filtering strategies we obtained the distribution of the lexical items that have been found in all samples and identified the five most frequent ones for each lexicon (Table 3). The comparison shows two different filtering patterns between the two strategies. The most-frequent words from Shutterstack lexicon focus on pornography (e.g., 'sex'); HateBase terms on racism (e.g., 'slaves', 'blacks') and misogyny (e.g., 'dykes'). Choosing one of the two lexicons does not have only an impact in the number of flagged sentences but also on the specific subset of potentially harmful contents that are removed.

[9]https://github.com/t-davidson/hate-speech-and-offensive-language/tree/master/lexicons

[10]https://pypi.org/project/alt-profanity-check/

[11]https://github.com/allenai/dolma/blob/main/python/dolma/\\taggers/jigsaw.py

[12]The People Dataset and all the implemented filtered strategies will be released under MIT license after the anonymity period.

lexicon	top-5
shutterstack	dick, sex, porn, ass, nude
hatebase	slave, married to, blacks, dykes, of white

Table 3: Top-5 matching lexical items in CC samples

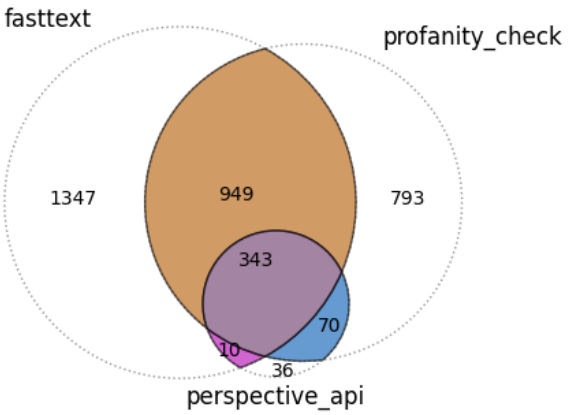

Figure 1: Intersection of contents removed through classifier-based strategies

The comparison of classifier-based strategies focuses on the number of flagged sentences in which their classification overlaps and those where it does not. In Figure 1, we can identify two patterns. The total number of sentences classified as harmful by Perspective API is almost a subset of ones classified by Profanity Check with an overlap of 90%. The overlap of Perspective API with the FastText classifier is lower but still significant (77%). This pattern can be explained by the small number of sentences flagged by Perspective API (459) compared to Profanity Check ($2,155$) and FastText ($2,649$). The second pattern is the strong difference in the sentences identified by Profanity Check and FastText: only 35.8% of sentences classified as harmful by FastText are also classified as such by Profanity Check. This shows that, similarly to the lexicon-based strategies, choosing one of the existing classifiers for harm detection may imply targeting different types of harmfulness.

Quality-based filtering strategies do not directly aim at harm detection but their training over allegedly high-quality documents is supposed to have an impact on the removal of toxic contents from raw documents. We check their effectiveness by counting the percentage of mentions kept by these strategies despite having been flagged as harmful by rule-based and classifier-based approaches. Results in Table 4 show that a high percentage of sentences classified as harmful are still present after the quality-based filtering. Adopting the classifier trained on Wikipedia documents leads to the removal of 15% of sentences while keeping 93.5% of harmful sentences. The classifier based on OpenWebText shows a similar proportion: it removes 45% of sentences but 69.3% of sentences classified as harmful by other strategies are still present after the filtering. This comparison

strategy	n. sents (%)	toxic sents (%)
wikipedia	85%	90.5%
webtext	55%	69.3%

Table 4: The percentage of harmful contents that remain after the application of quality-based strategies

reveals that quality is not a proxy of safety, since the most part of toxic contents is not removed through the adoption of quality-based strategies.

Who gets filtered out? A third line of analysis studies what entities are most impacted by filtering strategies. As can be seen in Table 2, the adoption of different strategies not only differs in its magnitude but systematically penalizes certain categories of people. Women are always the most impacted target in rule-based and classifier-based strategies, and in 4 cases out of 5, named entities that suffer the highest content removal are Western women. Conversely, strategies based on document similarity have diametrically opposed effect, since they impact the most on men. However, as shown in Table 4, filtering based on document quality is not a reliable method of identifying toxic contents. Therefore, it is not possible to assume that these strategies remove entity mentions that appear in harmful contents and their major impact on men cannot be interpreted as a side-effect of harm detection.

In order to deepen our analysis of the impact of data filtering strategies against groups, we leveraged our People Dataset to identify which are the occupations of named entities that are removed through data filtering strategies. For each entity mention occurring in a sentence that has been flagged as harmful by a rule-based or a classifier-based strategy we retrieved their occupation from our dataset and computed the occupations that occurred the most for each analyzed group. Since we observed that quality-based filtering strategies are not effective in the detection of toxic content, we did not consider them in this analysis.

Table 5 shows the 5 most occurring occupations in datasets and the 5 that have been mostly flagged for removal by filtering strategies. The analysis confirms the presence of patterns of discrimination along the gender axis. The distribution of men's occupations that have been filtered is coherent with the original distribution of mentions; for women this is not the same. The most removed occupation of Western women is 'pornographic actor', which is not among the most frequent occupations of Western women. Similarly, the profession 'model' is one of the most removed from postcolonial women occupations despite not being frequent in the original distribution. This means that the removal of specific forms of harm, which has not necessarily a negative effect, increases the underrepresentation of very specific categories of people, suggesting the need of adopting finer-grained analysis of the impact of filtering strategies on people.

group	top-k occupation	top-k filtered
w.m.	writer, politician, actor, film actor, television actor	actor, writer, film actor, politician, television actor
p-c. m.	politician, actor, writer, television actor, singer	politician, actor, singer, writer, film actor
w. w.	actor, film actor, singer, television actor, writer	pornographic actor, actor, film actor, film director, television actor
p-c. w.	actor, politician, writer, film actor, singer	actor, film actor, singer, model, writer

Table 5: top-5 removal by profession

5 Discussion

Our survey of data filtering strategies shows that the implementation of effective approaches to detect and remove harmful contents from pretraining datasets is still an open issue with important social implication. Choosing a strategy means addressing a specific subcategory of harmfulness and this has an impact against specific groups of people (RQ 2). Despite this variety, a general pattern emerges from the implementation of all rule-based and classifier-based strategies: **the systematic increase of underrepresentation of women in pretraining datasets**. Mentions of women are always removed at a highest rate than men. This systematic bias shows that simply removing data might not be an effective strategy to ensure LLM fairness. Alternative approaches can rely on data filtering strategies to guide the implementation of human-in-the-loop approaches (Ali et al. 2025) to the curation of pretraining datasets.

In contrast with this evidence about the complex nature of data filtering for harm reduction, the general tendency that emerges from our survey (RQ 1) is disengagement with this issue. After a wave of efforts in this field connected with the critical work of (Bender et al. 2021), **the interest in implementing robust pretraining data filtering dramatically reduced** in favor of post-training measures for harm reduction. In most cases, the actual implementation of strategies is close-sourced, hindering the replication of procedures that are adopted for processing datasets. The presence of notable exceptions such as Dolma (Soldaini et al. 2024) and the RefinedWeb Dataset (Penedo et al. 2023) represents a significant counter-tendency. However, there is still a lack of filtering strategies that are effective in balancing the need to remove harmful contents while preserving the representativeness of vulnerable groups in pretraining datasets.

6 Conclusion

In this paper we presented a first systematic analysis of data filtering strategies for harm reduction in datasets. The overview of 55 technical reports describing LLMs enabled us to draw a taxonomy of filtering strategies and to identify the open issues that prevent the implementation of effective, reliable, and replicable methods for the reduction of harmful contents. Additionally, we experimentally evaluated seven existing strategies on CC samples of documents in order to assess their impact against groups characterized by the intersection of gender and origin. The evaluation showed a systematic negative effect of strategies against women and a high variety in the types of harmful contents that specific strategies filter out.

Future work will focus on providing a more effective pipeline for the assessment of data filtering strategies with three aims in mind: *i.* improving the coverage of our pipeline by including additional information in the People Dataset. We plan to include information from additional knowledge bases and add information about vulnerable groups that goes beyond named entities (e.g., mentions of communities, ethnic groups, and demonyms) ; *ii.* adopting a participatory approach to the evaluation of filtering strategies that engages associations and communities of people who are active against discrimination. This approach will enable to explore the data filtering process from the perspectives of people who might actually harmed by the implementation of strategies; *iii* we will systematically explore the correlation between the adoption of specific data filtering strategies and linguistic behaviors emerging from models trained on filtered datasets.

Limitation

A first limitation of our research is the imbalance of the two knowledge bases that we use to obtain sociodemographic information about named entities. Both Wikipedia and Caligraph contain a higher number of Westerners against Postcolonial people and a higher number of men against women. This introduces a bias in the Entity Linking process that might have an impact on results. Developing a more balanced dataset for the implementation of our pipeline is one of the key actions of our future work.

A second limitation regards the focus on named entities. People belonging to vulnerable groups are mentioned in documents in very different ways (e.g., demonym, pronouns, countries of origin). However, there are no systems that are trained to automatically identify these triggers of vulnerable identities. We preferred to rely on state of the art approaches to maximize the precision of our EL pipeline against the recall. The second version of our pipeline will account for these alternative ways of mentioning people with certain sociodemographic characteristics.

Ethical Statement

Since our work focuses on vulnerable groups to discrimination, we are aware of the risk of adopting research design biases that can have a negative influence on our categorization of people. We followed the theoretical background emerging from post-colonial and black studies, to avoid the risk of inducing stereotypical representation of vulnerable groups in our analysis. We also acknowledge that the underrepresentation of post-colonial people in existing knowledge bases can be a source of additional discrimination that will be challenged in the next iterations of our benchmarking pipeline.

WARNING: the paper could contain racist, sexist, violent, and generally offensive contents

References

Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A. A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Albalak, A.; Elazar, Y.; Xie, S. M.; Longpre, S.; Lambert, N.; Wang, X.; Muennighoff, N.; Hou, B.; Pan, L.; Jeong, H.; et al. 2024. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*.

Ali, M.; Brack, M.; Lübbering, M.; Wendt, E.; Khan, A. G.; Rutmann, R.; Jude, A.; Kraus, M.; Weber, A. A.; Stollenwerk, F.; Kaczér, D.; Mai, F.; Flek, L.; Sifa, R.; Flores-Herr, N.; Koehler, J.; Schramowski, P.; Fromm, M.; and Kersting, K. 2025. Judging Quality Across Languages: A Multilingual Approach to Pretraining Data Filtering with Language Models. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 8870–8909. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Antropic. 2023. Model Card and Evaluations for Claude Models. https://www-cdn.anthropic.com/files/4zrzovbb/website/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226.pdf.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bandy, J.; and Vincent, N. 2021. Addressing" documentation debt" in machine learning: A retrospective datasheet for bookcorpus. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhang, H.; Zhu, B.; Jordan, M.; Gonzalez, J. E.; et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113.

Crenshaw, K. W. 2013. Mapping the margins: Intersectionality, identity politics, and violence against women of color. In *The public nature of private violence*, 93–118. Routledge.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, 512–515.

DeepSeek-AI; Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Guo, D.; Yang, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Dai, F.; Luo, F.; Hao, G.; Chen, G.; Li, G.; Zhang, H.; Bao, H.; Xu, H.; Wang, H.; Zhang, H.; Ding, H.; Xin, H.; Gao, H.; Li, H.; Qu, H.; Cai, J. L.; Liang, J.; Guo, J.; Ni, J.; Li, J.; Wang, J.; Chen, J.; Chen, J.; Yuan, J.; Qiu, J.; Li, J.; Song, J.; Dong, K.; Hu, K.; Gao, K.; Guan, K.; Huang, K.; Yu, K.; Wang, L.; Zhang, L.; Xu, L.; Xia, L.; Zhao, L.; Wang, L.; Zhang, L.; Li, M.; Wang, M.; Zhang, M.; Zhang, M.; Tang, M.; Li, M.; Tian, N.; Huang, P.; Wang, P.; Zhang, P.; Wang, Q.; Zhu, Q.; Chen, Q.; Du, Q.; Chen, R. J.; Jin, R. L.; Ge, R.; Zhang, R.; Pan, R.; Wang, R.; Xu, R.; Zhang, R.; Chen, R.; Li, S. S.; Lu, S.; Zhou, S.; Chen, S.; Wu, S.; Ye, S.; Ye, S.; Ma, S.; Wang, S.; Zhou, S.; Yu, S.; Zhou, S.; Pan, S.; Wang, T.; Yun, T.; Pei, T.; Sun, T.; Xiao, W. L.; Zeng, W.; Zhao, W.; An, W.; Liu, W.; Liang, W.; Gao, W.; Yu, W.; Zhang, W.; Li, X. Q.; Jin, X.; Wang, X.; Bi, X.; Liu, X.; Wang, X.; Shen, X.; Chen, X.; Zhang, X.; Chen, X.; Nie, X.; Sun, X.; Wang, X.; Cheng, X.; Liu, X.; Xie, X.; Liu, X.; Yu, X.; Song, X.; Shan, X.; Zhou, X.; Yang, X.; Li, X.; Su, X.; Lin, X.; Li, Y. K.; Wang, Y. Q.; Wei, Y. X.; Zhu, Y. X.; Zhang, Y.; Xu, Y.; Xu, Y.; Huang, Y.; Li, Y.; Zhao, Y.; Sun, Y.; Li, Y.; Wang, Y.; Yu, Y.; Zheng, Y.; Zhang, Y.; Shi, Y.; Xiong, Y.; He, Y.; Tang, Y.; Piao, Y.; Wang, Y.; Tan, Y.; Ma, Y.; Liu, Y.; Guo, Y.; Wu, Y.; Ou, Y.; Zhu, Y.; Wang, Y.; Gong, Y.; Zou, Y.; He, Y.; Zha, Y.; Xiong, Y.; Ma, Y.; Yan, Y.; Luo, Y.; You, Y.; Liu, Y.; Zhou, Y.; Wu, Z. F.; Ren, Z. Z.; Ren, Z.; Sha, Z.; Fu, Z.; Xu, Z.; Huang, Z.; Zhang, Z.; Xie, Z.; Zhang, Z.; Hao, Z.; Gou, Z.; Ma, Z.; Yan, Z.; Shao, Z.; Xu, Z.; Wu, Z.; Zhang, Z.; Li, Z.; Gu, Z.; Zhu, Z.; Liu, Z.; Li, Z.; Xie, Z.; Song, Z.; Gao, Z.; and Pan, Z. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Dodge, J.; Sap, M.; Marasović, A.; Agnew, W.; Ilharco, G.; Groeneveld, D.; Mitchell, M.; and Gardner, M. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1286–1305.

Du, N.; Huang, Y.; Dai, A. M.; Tong, S.; Lepikhin, D.; Xu, Y.; Krikun, M.; Zhou, Y.; Yu, A. W.; Firat, O.; et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, 5547–5569. PMLR.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dutta, A.; Khorramrouz, A.; Dutta, S.; and KhudaBukhsh, A. R. 2024. Down the toxicity rabbit hole: A framework to bias audit large language models with key emphasis on racism, antisemitism, and misogyny. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI*, 3–9.

Edwards, L. 2021. The EU AI Act: a summary of its significance and scope. *Artificial Intelligence (the EU AI Act)*, 1.

Erxleben, F.; Günther, M.; Krötzsch, M.; Mendez, J.; and Vrandečić, D. 2014. Introducing wikidata to the linked data web. In *The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I 13*, 50–65. Springer.

Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–3369.

Gemini, T.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gokaslan, A.; Cohen, V.; Pavlick, E.; and Tellex, S. 2019. OpenWebText Corpus. http://Skylion007.github.io/OpenWebTextCorpus.

Groeneveld, D.; Beltagy, I.; Walsh, E.; Bhagia, A.; Kinney, R.; Tafjord, O.; Jha, A.; Ivison, H.; Magnusson, I.; Wang, Y.; et al. 2024. OLMo: Accelerating the Science of Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15789–15809.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*.

Heist, N.; and Paulheim, H. 2019. Uncovering the semantics of Wikipedia categories. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I 18*, 219–236. Springer.

Intelligence, A. A. G. 2024. The Amazon Nova Family of Models: Technical Report and Model Card. https://assets.amazon.science/.

Jain, N.; Vaidyanath, S.; Iyer, A.; Natarajan, N.; Parthasarathy, S.; Rajamani, S.; and Sharma, R. 2022. Jigsaw: Large Language Models meet Program Synthesis. In *ICSE 2022*.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jo, E. S.; and Gebru, T. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 306–316.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, 2. Minneapolis, Minnesota.

Laurençon, H.; Saulnier, L.; Wang, T.; Akiki, C.; Villanova del Moral, A.; Le Scao, T.; Von Werra, L.; Mou, C.; González Ponferrada, E.; Nguyen, H.; et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35: 31809–31826.

Lieber, O.; Lenz, B.; Bata, H.; Cohen, G.; Osin, J.; Dalmedigos, I.; Safahi, E.; Meirom, S.; Belinkov, Y.; Shalev-Shwartz, S.; et al. 2024. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*.

Liu, Y. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Longpre, S.; Yauney, G.; Reif, E.; Lee, K.; Roberts, A.; Zoph, B.; Zhou, D.; Wei, J.; Robinson, K.; Mimno, D.; et al. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3245–3276.

Luccioni, A. S.; and Viviano, J. D. 2021. What's in the box? a preliminary analysis of undesirable content in the common crawl corpus. *arXiv preprint arXiv:2105.02732*.

Lucy, L.; Gururangan, S.; Soldaini, L.; Strubell, E.; Bamman, D.; Klein, L. F.; and Dodge, J. 2024. AboutMe: Using self-descriptions in webpages to document the effects of english pretraining data filters. *arXiv preprint arXiv:2401.06408*.

Manghi, P. 2023. Miriam Baglioni1, Andrea Mannocci1, Gina Pavone1, Michele De Bonis1 and. 47–59.

Mickus, T.; Grönroos, S.-A.; Attieh, J.; Boggia, M.; de Gibert, O.; Ji, S.; Loppi, N. A.; Raganato, A.; Vázquez, R.; and Tiedemann, J. 2024. MAMMOTH: Massively Multilingual Modular Open Translation@ Helsinki. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 127–136.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

Penedo, G.; Malartic, Q.; Hesslow, D.; Cojocaru, R.; Cappelli, A.; Alobeidli, H.; Pannier, B.; Almazrouei, E.; and Launay, J. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Rae, J. W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; Young, S.; et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

Research, L.; An, S.; Bae, K.; Choi, E.; Choi, K.; Choi, S. J.; Hong, S.; Hwang, J.; Jeon, H.; Jo, G. J.; et al. 2024. EXAONE 3.5: Series of Large Language Models for Real-world Use Cases. *arXiv preprint arXiv:2412.04862*.

Research, M. 2023. Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs. https://www.databricks.com/blog/mpt-7b.

Soldaini, L.; Kinney, R.; Bhagia, A.; Schwenk, D.; Atkinson, D.; Authur, R.; Bogin, B.; Chandu, K.; Dumas, J.; Elazar, Y.; et al. 2024. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15725–15788.

St, L.; Wold, S.; et al. 1989. Analysis of variance (ANOVA). *Chemometrics and intelligent laboratory systems*, 6(4): 259–272.

Stranisci, M. A.; Bernasconi, E.; Patti, V.; Ferilli, S.; Ceriani, M.; and Damiano, R. 2023a. The world literature knowledge graph. In *International Semantic Web Conference*, 435–452. Springer.

Stranisci, M. A.; Damiano, R.; Mensa, E.; Patti, V.; Radicioni, D.; and Caselli, T. 2023b. WikiBio: a Semantic Resource for the Intersectional Analysis of Biographical Events. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12370–12384.

Su, G.; Wu, M.; Huang, Z.; Zhang, Y.; Wang, T.; Hu, Y.; and Sha, Y. 2024. Refine, align, and aggregate: multi-view linguistic features enhancement for aspect sentiment triplet extraction. In *Findings of the Association for Computational Linguistics ACL 2024*, 3212–3228.

Sun, Y.; Wang, S.; Feng, S.; Ding, S.; Pang, C.; Shang, J.; Liu, J.; Chen, X.; Zhao, Y.; Lu, Y.; et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

team, G. 2024. Grok-2 Beta Release. https://x.ai/blog/grok-2.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; et al. 2024a. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Team, R.; Ormazabal, A.; Zheng, C.; d'Autume, C. d. M.; Yogatama, D.; Fu, D.; Ong, D.; Chen, E.; Lamprecht, E.; Pham, H.; et al. 2024b. Reka core, flash, and edge: A series of powerful multimodal language models. *arXiv preprint arXiv:2404.12387*.

Tedeschi, S.; Friedrich, F.; Schramowski, P.; Kersting, K.; Navigli, R.; Nguyen, H.; and Li, B. 2024. ALERT: A Comprehensive Benchmark for Assessing Large Language Models' Safety through Red Teaming. *arXiv preprint arXiv:2404.08676*.

Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tunstall, L.; Beeching, E.; Lambert, N.; Rajani, N.; Rasul, K.; Belkada, Y.; Huang, S.; von Werra, L.; Fourrier, C.; Habib, N.; et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Wake, A.; Wang, A.; Chen, B.; Lv, C.; Li, C.; Huang, C.; Cai, C.; Zheng, C.; Cooper, D.; Dai, E.; et al. 2024. Yi-lightning technical report. *arXiv preprint arXiv:2412.01253*.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Weber, M.; Fu, D.; Anthony, Q.; Oren, Y.; Adams, S.; Alexandrov, A.; Lyu, X.; Nguyen, H.; Yao, X.; Adams, V.; et al. 2024. Redpajama: an open dataset for training large language models. *arXiv preprint arXiv:2411.12372*.

Xie, S. M.; Santurkar, S.; Ma, T.; and Liang, P. S. 2023. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems*, 36: 34201–34227.

Xu, A.; Pathak, E.; Wallace, E.; Gururangan, S.; Sap, M.; and Klein, D. 2021. Detoxifying Language Models Risks Marginalizing Minority Voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2390–2397.

Yang, Z. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint arXiv:1906.08237*.

Yasunaga, M.; Leskovec, J.; and Liang, P. 2022. LinkBERT: Pretraining Language Models with Document Links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8003–8016.

Zhong, Q.; Ding, L.; Zhan, Y.; Qiao, Y.; Wen, Y.; Shen, L.; Liu, J.; Yu, B.; Du, B.; Chen, Y.; et al. 2022. Toward efficient language model pretraining and downstream adaptation via self-evolution: A case study on superglue. *arXiv preprint arXiv:2212.01853*.

Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 19–27.