The Promises and Pitfalls of LLM Annotations in Dataset Labeling: a Case Study on Media Bias Detection

Tomáš Horych[1], Christoph Mandl[1], Terry Ruas[1], André Greiner-Petter[2], Bela Gipp[1], Akiko Aizawa[2], Timo Spinde[1]
[1]University of Göttingen, Göttingen, Germany
[2]National Institute of Informatics, Tokyo, Japan
t.horych@media-bias-research.org,
c.mandl@media-bias-research.org, ruas@uni-goettingen.de,
greinerpetter@gipplab.org, gipp@uni-goettingen.de,
aizawa@nii.ac.jp, t.spinde@media-bias-research.org

Abstract

High annotation costs from hiring or crowdsourcing complicate the creation of large, high-quality datasets needed for training reliable text classifiers. Recent research suggests using Large Language Models (LLMs) to automate the annotation process, reducing these costs while maintaining data quality. LLMs have shown promising results in annotating downstream tasks like hate speech detection and political framing. Building on the success in these areas, this study investigates whether LLMs are viable for annotating a complex task of media bias detection and whether a downstream media bias classifier can be trained on such data. We create *Anno-lexical* , the first large-scale dataset for media bias classification with over 48k synthetically annotated examples. Our classifier fine-tuned on it surpasses all of the annotator LLMs by 5-9% in Mathew's Correlation Coefficient (MCC) and performs close to or outperforms the model trained on human-labeled data when evaluated on two media bias benchmark datasets (BABE and BASIL). This study demonstrates how our approach significantly reduces the cost of dataset creation in the media bias domain and, by extension - the development of the classifiers, while our subsequent behavioral stress-testing reveals some of its current limitations and trade-offs.

1 Introduction

Media bias detection requires high-quality annotations to train classifiers that accurately identify biases across the political spectrum (Wessel et al., 2023; Spinde et al., 2021b). Cognitive biases and limited experience often make it hard for raters to annotate bias accurately (Spinde et al., 2021a), leading to inconsistent annotations across annotators and instances (Spinde et al., 2021c). Achieving such high-quality annotations is challenging due to the resource-intensive nature of the task and the need for domain expertise (Monarch, 2021). Popular expert-based datasets in this domain, like BABE (Spinde et al., 2021c) and BASIL (Fan et al., 2019), contain a limited number of labeled sentences (4k and 10k, respectively), with the need for experts and the associated costs limiting their size. This limitation, in turn, affects the performance of the resulting models (Spinde et al., 2021c). The difficulties in creating datasets also affect dataset diversity, which is crucial when improving media bias classification performances (Horych et al., 2024).

Although crowdsourcing is a viable approach to scale data annotation, crowdsource workers often do not have sufficient experience to judge bias correctly (Spinde et al., 2021c). Even more, the quality of crowdsourced labels, particularly from major platforms like Amazon MTurk, has significantly declined over the years (Chmielewski and Kucker, 2020a). This decline is a common problem in media bias detection and other areas of machine learning and natural language processing (NLP) (Chmielewski and Kucker, 2020a). LLMs offer promising opportunities to support human annotators by automating the annotation process, ensuring consistency, and adapting to specific domains, which can reduce costs and improve or sustain quality. (Gilardi et al., 2023; Alizadeh et al., 2023; He et al., 2024; Tan et al., 2024). However, while current research focuses on evaluating LLMs' general capabilities on NLP benchmarks, the viability of learning from LLM-made annotations in com-

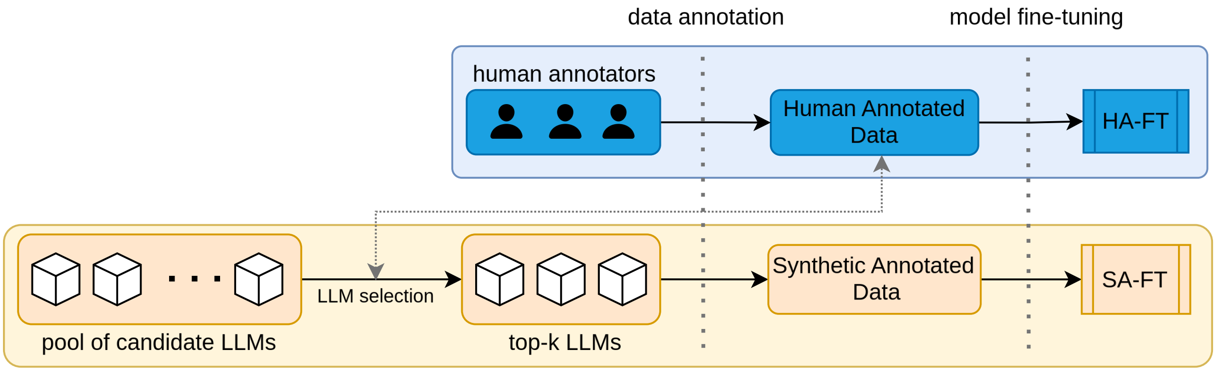

Figure 1: Workflow diagram presenting the difference between the two approaches to fine-tuning the model - **H**uman-**A**nnotation **F**tine-**T**uning (**HA-FT**) and **S**ynthetic-**A**nnotation **F**tine-**T**uning (**SA-FT**). The grey arrow between "LLM selection" and "Human Annotated Data" represents an optional step for informed LLM selection.

plex downstream tasks like media bias detection remains underexplored.

In this work, we investigate whether LLMs can provide annotations of sufficient quality to be used to train smaller models for the particular classification task of media bias classification. We pick a lexical bias classification as a focal task, as its reliance on the lexical features (see Section 2) makes it the most popular subtask of a general media bias classification task among media bias researchers in the NLP domain (see Section 3). For a detailed overview of media bias and how the components can be defined, we refer to the literature reviews by Rodrigo-Ginés et al. (2024) and Spinde et al. (2023). We introduce a three-stage pipeline to analyze the feasibility of learning lexical bias detection from LLM annotations. We select three LLMs based on an a priori evaluation, and with a few-shot in-context learning prompt, we label a large-scale training dataset - *Anno-lexical* . Finally, we fine-tune a classifier on the aggregated majority-vote label of the *Anno-lexical* dataset. This approach and its comparison to conventional fine-tuning is depicted in the Figure 1.

We compare a Synthetic-Annotations Fine-Tuned classifier (**SA-FT**) against the conventional classifier fine-tuned on human annotations (**HA-FT**) in terms of both performance and robustness. Our study answers the following research question:

- **RQ1**: Can SA-FT match the performance of the HA-FT on state-of-the-art lexical bias benchmarks?
- **RQ2**: Is the SA-FT classifier robust against spurious correlations?

The contributions of our work are as follows:

- We show that the SA-FT classifier outperforms their teacher LLMs and performs comparably with the conventional HA-FT on the sentence-level lexical bias classification task.
- We show that the SA-FT classifier's performance stems from its strength in recalling a major portion of the positive class, but its precision and robustness to input perturbations are worse than that of HA-FT.
- We publish the *Anno-lexical* dataset, a large-scale dataset with 48330 sentences with synthetic lexical bias annotations.

Additionally, we publish a Python package named *Annomatic* simplifying the annotation pipeline with LLMs. The code, data, and trained SA-FT classifier are publicly available at:

github.com/Media-Bias-Group/llm-annotations-annomatic

2 Focal task definition

In this work, we focus on the binary classification of lexical bias at the sentence level. According to Fan et al. (2019), lexical bias stems from the choice of words and can be identified based solely on lexical features within a sentence. We use a definition of lexical bias instead of linguistic bias as the latter sometimes refers only to morphological text aspects and is generally used with less consistency (Spinde et al., 2023). In this work, we will refer to lexical bias as the narrow focal task and media bias as the superset of the bias detection tasks.

3 Related Work

Sentence-Level Lexical Bias Detection. Only a few dedicated human-labeled sentence-level lexical bias detection datasets exist, such as MBIC (1700 sentences) (Spinde et al., 2021d), BASIL (7919 sentences) (Fan et al., 2019) and BABE (4121 sentences) (Spinde et al., 2021c). Given the diversity of language and the multitude of options to portray content, especially the small size limits the performance of lexical bias classifiers, failing to capture its diverse manifestation (Wessel and Horych, 2024). Methods for sentence-level bias detection often involve fine-tuning pre-trained language models on these datasets. To address the scarcity of ground truth data, researchers have explored various transfer learning strategies, including distant supervision (Spinde et al., 2021c), event relation graph augmentation (Lei and Huang, 2024), domain-adaptive pre-training (Krieger et al., 2022), fine-grained bias indicators (Lin et al., 2024), and multi-task learning (Spinde et al., 2022; Horych et al., 2024). These transfer learning approaches have consistently yielded positive results, showing the benefits of dataset diversity in the domain. The top-performing MAGPIE model achieves an F1 score of 0.841 (Horych et al., 2024) on BABE. However, these methods address the lack of high-quality training data only indirectly (transfer learning and data augmentation). The main issue—relying on expert-labeled data—remains, as obtaining these labels is both time-consuming and expensive, which limits the necessary scale. This study aims to directly address the problem of sourcing primary data.

LLM dataset labeling. Traditional annotation methods face high costs and quality issues (Klie et al., 2023; Marshall et al., 2023; Chmielewski and Kucker, 2020b). Advances in LLMs suggest they can be efficient alternatives at considerably lower costs. Studies show LLMs can match or exceed human annotators in tasks like implicit hate speech detection (Törnberg, 2023; Huang et al., 2023; He et al., 2024) and political framing detection (Gilardi et al., 2023; Alizadeh et al., 2023). While much research on the topic focuses on ChatGPT (which mostly also shows superior annotation quality across tasks), evaluating open-source models for tasks like lexical bias detection is crucial to ensure broader accessibility and cost-effectiveness in NLP applications (Gilardi et al., 2023; Alizadeh et al., 2023; He et al., 2024). Across models, experiments demonstrate that few-shot approaches, as well as techniques like Chain-of-Thought (CoT) (Wei et al., 2022) and explanatory methods, significantly improve annotation quality (Gilardi et al., 2023; Alizadeh et al., 2023; He et al., 2024). We discuss in Section 7 how human input and evaluation will, therefore, still remain crucial for achieving the best results in any approach, including automated annotations.

4 Annotations with LLMs

This section describes the process of the synthetic annotation. Since reproducibility is a significant challenge in the NLP domain (Belz et al., 2021, 2023), we developed *Annomatic* - a robust tool to make our experiments easily reproducible. The principal objectives of *Annomatic* are to 1) abstract away the setup of LLMs from different sources, 2) parse & interpret the LLM output, and 3) aggregate the results of multiple LLMs in an ensemble. We provide more detail about the *Annomatic* in the code repository[1] and in the Appendix B.

4.1 Annotation workflow

We employ general-purpose LLMs (e.g., LLama2) as annotators, which annotate in a scenario that we refer to as *near-unsupervised*. In this approach, the LLMs generate off-the-shelf annotations with minimal direct human intervention. The only human signal (supervision) provided comes from a set of human-labeled examples included in the prompts, used to guide the model's in-context learning. We elaborate on the constraint of *near-unsupervision* in Section 7.

The annotation process begins with prompting the annotator LLMs. We use a few-shot in-context learning format to prompt the LLM. The prompt consists of the following components:

Examples - up to 8 examples of human-labeled sentences from a pool of 100 selected sentences. Specifically, we randomly sampled 100 human-labeled examples from the BABE dataset (Spinde et al., 2021c), the ground truth dataset for the lexical bias detection task.

Explanations - alongside each example, an explanation generated by GPT-4 (OpenAI et al., 2024) is provided. The explanation is a short text describing how the label in the example was determined.

Target of annotation - the last component is a target sentence to be annotated and a short instruction

[1] https://github.com/Media-Bias-Group/annomatic/

with label options (e.g., "Contains lexical bias" / "Does not contain lexical bias"). Table 8 (Appendix) contains the full prompt template used.

The examples and explanations are selected for each sentence instance individually. In the annotation (inference) time, we retrieve the *k* most similar labeled examples for each target sentence using the KATE algorithm (Liu et al., 2022) and using the similarity measure as a retrieval criterion, based on the findings of (Margatina et al., 2023). Once the LLMs have processed all the data, we parse the responses to extract the final label. We search for the most frequently occurring label in the response and match them against a list of *positive* and *negative* label options manually curated by the authors. If no labels appear in the response or the labels result in a tie, we label it with a question mark '?'. We later manually review these ambiguous cases and exclude sentences with inconclusive responses. Finally, a common practice in a human-annotator setting is determining labels through majority voting among the human annotators. In an attempt to mitigate any potential cultural and social biases of LLMs (Liang et al., 2021; Navigli et al., 2023), we adopt the same approach by inferring the final label through majority voting among the LLM annotators.

In addition to the open-source code on Github, we release our annotation tool on PyPi[2] under Apache-2.0 license to reduce efforts to replicate our work and simplify its adoption in new projects.

4.2 Annotator selection

To select the LLM annotators, we evaluate a pool of open- and closed-source models on the training set of the BABE dataset. The goal of this evaluation is two-fold: to verify that LLMs without fine-tuning can detect lexical bias, thereby qualifying as annotators, and to construct a ranking that we use to select the final annotators. The candidate LLMs are selected based on the snapshot of the Open LLM leaderboard[3] at the time of the experiments. We chose seven open-source general-purpose LLMs from the top of the leaderboard: Falcon 7B Instruct, Zephyr 7B beta, OpenChat 3.5, Mistral-7B-v0.1 Instruct, Mistral-8x7B instruct, LLama 2 7B, 13B and two closed-source models GPT-4-turbo and GPT-3.5-turbo, to cover the closed-source state-of-the-art. Additionally, we include four models from the FLAN encoder-decoder model family (Raffel et al., 2020) in sizes ranging from Base to Ultra-Large due to their demonstrated effectiveness in classification tasks (Ziems et al., 2024). The list of all models, together with references and basic information, can be found in the Appendix 4. The evaluation results are presented in Table 1. While the proprietary GPT-4 outperforms every open-source model, three of the open-source models outperform GPT-3.5 in five prompting settings (marked with $*$). However, the performance of LLMs on the BABE dataset highlights that there remains room for improvement in the lexical-bias detection task, especially for open-source models. Due to the cost constraints, we exclude GPT-4, GPT-3.5, and Mixtral-8x7b from the final annotator selection. We fix the number of selected LLM annotators - k to three - the lowest odd number that will ensure a majority decision while keeping the cost efficiency. Using an odd number of models guarantees a clear majority label. Increasing k to five or more while potentially improving accuracy would result in higher computational costs. Therefore, based on the performance, Zephyr 7B Beta, OpenChat 3.5, and LLama 2 13B Chat are selected to annotate the downstream task in the setting with the highest mean performance: 8-shot explanation.

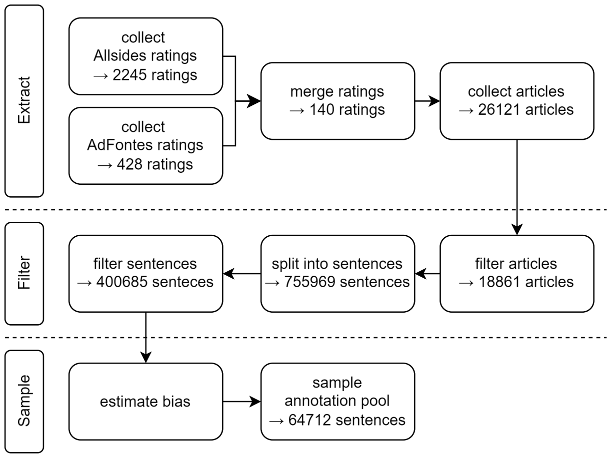

Figure 2: The workflow diagram describing an end-to-end construction of our politically balanced text corpus.

5 A synthetic bias classifier

This section presents our proposed process of developing a lexical bias classifier fine-tuned only on the SA-FT. The process consists of three steps: 1. Curating an annotation corpus, 2. Annotating the corpus, 3. Training a classifier on the synthetic annotations.

[2] https://pypi.org/
[3] huggingface/open-llm-leaderboard

model	0-shot	+ sys prompt	0-shot Exp	2-shot	4-shot	8-shot	2-shot Exp	4-shot Exp	8-shot Exp	mean
Zephyr 7B beta	**0.551***	0.385	0.369	0.538	0.548	0.558	0.6	0.616	0.627	0.532
OpenChat 3.5	0.389	0.499	**0.503**	0.577	0.581	**0.593***	0.565	0.58	0.622	**0.546**
Mistral-7B-v0.1 Instruct	0.343	0.357	0.248	0.353	0.415	0.46	0.487	0.495	0.534	0.41
LLama 2 7B Chat	0.15	0.101	0.294	0.359	0.416	0.497	0.554	0.581	0.579	0.392
LLama 2 13B Chat	0.238	0.032	0.325	0.406	0.448	0.517	0.619	0.619	0.613	0.424
Flan-UL2	0.489	**0.534**	0.462	0.532	0.526	0.537	0.432	0.459	0.516	0.499
Falcon-7B-Instruct	0.052	0.038	0.128	0.175	0.227	0.178	0.344	0.304	0.274	0.191
FLAN-T5-XL	0.302	0.356	0.346	0.406	0.415	-	-	-	-	0.365
FLAN-T5-Large	0.133	0.312	0.335	0.165	0.146	-	-	-	-	0.218
FLAN-T5-Base	0.107	0.12	0.061	0.044	0.044	-	-	-	-	0.075
Mixtral-8x7B Instruct	0.277	0.279	0.494	**0.583**	**0.595***	0.588	**0.646***	**0.654***	**0.662**	0.531
GPT-3.5 Turbo	0.511	0.596	0.56	0.595	0.586	0.591	0.624	0.633	0.663	0.595
GPT-4 Turbo	0.683	0.697	-	0.71	0.699	0.7	0.83	0.786	0.753	0.732
average	0.325	0.331	0.344	0.419	0.434	0.522	0.57	0.572	**0.584**	

Table 1: All results are measured with Mathew's Correlation Coefficient (MCC) on the BABE train/development (combined) set. Bold scores mark the best-performing open-source model for a given prompting. An asterisk * marks performance higher than GPT-3.5. The blank spots (−) mark runs where A) the size of the model's context window is insufficient and B) the model's output diverges from the instruction.

5.1 The annotation corpus

This section outlines the process of creating our unlabeled text corpus consisting of news sentences for the downstream annotation. Given the sensitive nature of bias detection, related work highlights the importance of well-balanced data sources (Scheuerman et al., 2021; Fan et al., 2019). Moreover, as LLMs exhibit inherent cultural and social biases (Liang et al., 2021; Navigli et al., 2023), an imbalance in the distribution of the political spectrum could lead to skewed models, amplifying existing biases rather than enabling their detection. Therefore, we propose a strict balancing procedure as a pre-emptive countermeasure to the potential biases.

We use the platforms allsides.com and adfontesmedia.com to assess the underlying political leaning of the news text in a left-to-right manner. Figure 2 presents a workflow diagram of the corpus construction. We break the process down into three parts:

Extract. We start with collecting all public articles from outlets that have ratings from both *allsides* and *adfontesmedia* platforms. Both platforms use left-to-right ratings with different scales; we unify their ratings into five labels: Left, Lean Left, Center, Lean Right, and Right. We only keep articles where both platforms agree on the rating.

Filter. We filter out empty, short, or other corrupted articles and keep only articles written in English. We then segment these articles into sentences. Additionally, we trim special characters and other irregularities from the sentences. The final collection of filtered sentences contains approximately 400,000 sentences.

Sample. Finally, We sample sentences to ensure the balance across the aforementioned political spectrum. However, we can't ensure a fair distribution of lexical bias before knowing the true labels (i.e., before annotation). Some outlets may be more likely to contain lexical bias, which could result in an uneven distribution, with one side of the spectrum having mostly biased sentences and the other side being largely neutral after the annotation.

To tackle this issue, we implement a *pre-classification* stage using a state-of-the-art lexical bias classifier (Horych et al., 2024) to **estimate** the sentence's lexical bias before the annotation.

We use this prior bias estimate to sample sentences such that each segment of the political spectrum contains an equal number of sentences, with exactly 50% estimated to exhibit lexical bias and 50% exhibiting no bias. This downsampling leads to 64,712 sentences. This procedure helps to achieve a roughly equal lexical bias distribution across the political spectrum before the costly annotation.

By estimating the bias and downsampling based on that estimate, we prevent potentially large discarding of sentences after the annotation, given that Horych (2022) found that, on average, only 10% of sentences were biased in a sample of news articles.

5.2 Learning from synthetic annotations

Finally, an ensemble of the three chosen LLM annotators annotates the 64,712 sentences via majority vote (as it is usually done with human annotators). We use the majority vote instead of exploiting the best-performing model to make our synthetic annotation robust against potential model-specific fea-

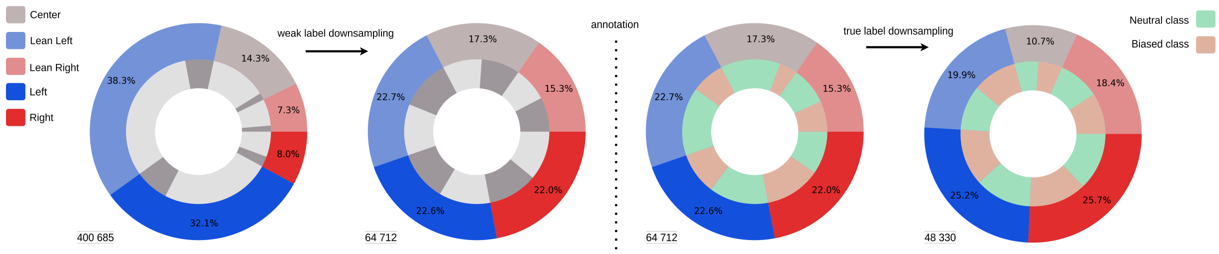

Figure 3: The figure demonstrates the transformation of the unlabeled corpus (left) to the final *Anno-lexical* annotated dataset (right) in terms of its size, political ideology distribution, and lexical bias distribution. The inner part of the pie charts represents the distribution of bias labels (neutral/biased) within each part of the political spectrum. The grey depiction of this distribution in the first two plots represents the weak labels estimated before annotation, and the colored (green and red) depiction represents the distribution of the true (annotated) labels.

tures and tendencies (Navigli et al., 2023; Liang et al., 2021). This results in 64,712 sentences annotated with lexical bias labels. We, however, continue to reduce the size of this dataset to ensure an exactly fair distribution of lexical bias labels among the segments of the political spectrum. For each spectrum segment, we again downsample the sentences, now based on the label obtained through annotation, in a $1:1$ ratio. The final version of the dataset contains 48,330 sentences. A diagram summarizing the transformation of the corpus's size and party/label distribution to the final dataset is presented in Figure 3.

We call this final dataset *Anno-lexical* , and we make it publicly available on our repository[1], pre-split into train/dev/test sets with a 0.7, 0.15, and 0.15 proportion, respectively. A sample of 60 sentences from the dataset, sampled for each poltiical leaning category, is shown in Appendix C.

As a last and final step, we fine-tune a RoBERTa[4] encoder LM with a 2-layer classification head on the *Anno-lexical* . As our work focuses on comparing two training data scenarios, we keep the model architecture constant to minimize its impact and do not experiment with more models. We refer to this model as a **SA-FT** classifier, and we put it to the test in the experiments in the following sections.

6 Experiments

In this section, we present the results of two evaluations of the SA-FT classifier to showcase its properties. First, we compare the performance of the SA-FT on two well-established lexical bias test sets - BABE and BASIL and compare it to the conventional HA-FT model (Section 6.3). Secondly, we stress-test the model with a dedicated adversarial test set - CheckList (Ribeiro et al., 2020), assessing its robustness against spurious cues and other shortcuts (Section 6.4).

6.1 Datasets

For the evaluation, we use two key datasets in the sentence-level lexical bias domain: **BABE** (Spinde et al., 2021c) - consists of 4121 sentences annotated for binary labels 0 (unbiased) and 1 (biased). **BASIL** (Fan et al., 2019) - consists of 7919 sentences annotated for ternary labels 0 (unbiased), 1 (lexical biased), 2 (informational-biased). We treat the lexical bias label as a positive class and the informational bias and unbiased as a negative class to unify the task with BABE and *Anno-lexical* .

6.2 Experimental setup

For all experiments, we report Matthew's Correlation Coefficient (MCC) as the primary evaluation metric for binary classification due to its higher robustness over the F1 score, as MCC provides a more balanced measure by considering all elements of the confusion matrix (Chicco and Jurman, 2020). We also report the F1 score to underline its shortcomings in the evaluation. For the BABE dataset, we use splits provided by the authors with 75% of training data and 25% of test data (1000 sentences). We use the entire training set of the BABE dataset to train the HA-FT model and to rank the LLM annotators, as described in Section 4.2. We then use the BABE test set for the evaluations. From the BASIL dataset, we use all 7919 sentences for the evaluations. We execute all experiments and annotations on one Nvidia A100 GPU. All training and evaluations were run as a single run.

[4] FacebookAI/roberta-base

	$size_{train}$	**$BABE_{test}$** P	R	F1	MCC	**BASIL** P	R	F1	MCC
Zephyr 7B beta	8-shot	0.831	0.773	0.801	0.569	-	-	-	-
OpenChat 3.5	8-shot	0.814	0.825	0.819	0.588	-	-	-	-
LLama 2 13B Chat	8-shot	0.828	0.834	0.831	0.614	-	-	-	-
majority vote	-	0.852	0.823	0.837	0.639	-	-	-	-
SA-FT	34k	0.875	0.814	0.843	0.662	**0.171**	0.502	**0.254**	**0.205**
HA-FT	3k	**0.916**	0.772	0.838	**0.678**	0.169	0.384	0.235	0.174
$SA\text{-}FT_{coreset}$	3k	0.829	**0.859**	**0.844**	0.638	0.136	**0.696**	0.228	0.201

Table 2: The results of the evaluation of the LLM annotators, SA-FT classifier, and HA-FT on two lexical bias benchmark test sets. The highest values within each column are marked in bold.

6.3 A downstream SA-FT generalization

In this experiment, we evaluate the generalization ability of the SA-FT classifier in three settings:

- **Comparison with the teacher models.** We evaluate the SA-FT classifier on the BABE test set against the three LLM annotators that annotated its training data.
- **Comparison with the HA-FT.** We compare the BABE test performance of SA-FT against HA-FT.
- **Performance on out-of-distribution test set.** Finally, we compare the SA-FT and the HA-FT classifiers on the held-out BASIL dataset.

The evaluation results are presented in Table 2. We report the following findings. First, we observe an improvement in the MCC metric of 2.3% in the SA-FT performance over the majority vote of the annotators and 5-9% compared to the single LLMs. We want to point out that the SA-FT classifier is smaller than the original annotators and was trained on their majority vote. This result demonstrates that our proposed framework achieves a 5% improvement over the best of the chosen LLM annotators (LLama 2 13B chat) while reducing the cost of deployment by a factor of 100[5] or 300 if a majority vote is used. We attribute the gap between the majority-voted label and the SA-FT performance to the achieved generalization from the synthetic annotations.

While the SA-FT classifier generalizes from the synthetic annotations, the HA-FT classifier, fine-tuned on the BABE training data, still outperforms the synthetic model by 1,5%. However, the HA-FT model has the advantage of being tested on data from the same distribution as its training set[6]. In other words, HA-FT model parameters are directly optimized on the BABE training set, whereas SA-FT is only optimized on the *Anno-lexical* training set. Therefore, we also evaluate both models on the held-out BASIL dataset, which neither of the models has seen in any way. In this evaluation, the SA-FT classifier outperforms the HA-FT model by 3.1%. We verify this result with the McNemar paired test for labeling disagreements (Gillick and Cox, 1989) and find it statistically significant (with $p<0.05$). However, we find that both models perform poorly, and the SA-FT model only recalls a slightly larger portion of the positive class. This low score can be partially attributed to a label transformation that we applied, where we consider an "information bias" class as a neutral class. We observed that both models often classify the information bias class as positive. Furthermore, it is important to note that these results represent a relative performance between SA-FT and HA-FT. Neither of the two methods represents state-of-the-art results on the BASIL dataset and may not be directly comparable to the previous results on the BASIL dataset due to our label transformation.

Finally, because *Anno-lexical* is larger (34k) than the BABE training set (3k), we create a coreset of the *Anno-lexical* , with the same size (3k) by following the approach of (Chai et al., 2023) and fine-tune another model on this coreset. We denote this model as $SA\text{-}FT_{coreset}$.

In the fair comparison regarding training size, we observe that the $SA\text{-}FT_{coreset}$ falls short and underperforms the HA-FT by 4% in MCC on the BABE test set. However, while it still outperforms HA-FT on the BASIL dataset in the MCC metric,

[5]The LLama 2 13B Chat has 13 billion parameters, while our RoBERTa SA-FT classifier has roughly 130 million.

[6]Both sets are two different splits of one dataset.

Test Type	Test	Examples	HA-FT	SA-FT
MFT	**1. factual test**: Short neutral sentences from fact-checking datasets should be unbiased.	"What is a stereotype? An unfair, generalization about a group of people."	0.951	**0.964**
INV	**1. locations test**: Replace locations with random locations should not change the label.	" Hawaii $\rightarrow$ U.S. eyes even stricter gun laws in wake of shooting that killed 2 police officers."	**0.984**	0.971
	2. pronouns test: Replacing named entities with He/She/Them should not change the label.	Despite Portman's $\rightarrow$ her insistence that she has tried to advance female directors, only one of her feature films was directed by a female.	**0.971**	0.957
	3. prejudice test: Replacing one minority with other minority should not change the label.	"For some people, Buddha $\rightarrow$ Christ holds immense significance."	**0.895**	0.852
DIR	**1. loaded-words test**: Injecting biased adjectives and biased adverbs should change neutral to biased.	"The EU has shockingly secured up to outrageous 400 million doses of AstraZeneca 's experimental vaccine."	0.413	**0.664**

Table 3: This table shows examples and results of the *CheckList* stress-testing of the two models. Model trained on human labels - **HA-FT** and the one trained on synthetic labels - **SA-FT**. The examples for each test represent instances where the model with the lower score on the right failed and the other succeeded. The formatting and style of this table are inspired by the tables used in the original *CheckList* paper (Ribeiro et al., 2020).

its performance is even more skewed to low precision and high recall. Compared with the full SA-FT version, the smaller SA-FT$_{\text{coreset}}$ model sacrifices precision for recall, whereas SA-FT's full-scale training learns to make better predictions from the synthetic data and achieves a higher MCC in both cases.

6.4 Robustness against shortcuts

While the SA-FT classifier can match the HA-FT model in raw performance, lexical bias detection is a subtle task, and the conventional performance metrics may not fully capture the model's behavior. In previous sections, we describe two pre-emptive countermeasures that we took to promote SA-FT downstream robustness: an LLM majority vote approach to data labeling in Section 4.1 and a strict balancing strategy in the training dataset construction in Section 5.1.

To quantify this robustness, we adapt the idea of *CheckList* - a behavioral stress-testing of the classifiers (Ribeiro et al., 2020) and extend its prior adoption in the media bias domain (Wessel and Horych, 2024).

Inspired by the original *CheckList*, we use three high-level tests: MFT (Minimum Functionality Test), INV (Invariance Test), and DIR (Directional Expectation Test). Please refer to the work by Ribeiro et al. (2020) for further information about the *CheckList* method.

We then use the *CheckList* to again, compare the SA-FT and HA-FT classifiers. The full description of each test case with examples and the results of each model are presented in Table 3.

The SA-FT demonstrates a minor advantage (1.4%) in the minimum functionality test and a significant advantage (20+%) in the directional expectation test, where we expected the introduction of loaded words to change the neutral label to positive. In other words, the SA-FT is more attentive to strongly connotated words (e.g., *shockingly*, *terrible*). We also argue that these results align with generally lower recall of the HA-FT classifier in Section 6.3. However, the HA-FT prevails on every invariance test, which tests the models' sensitivity to input perturbance. These results show that while the SA-FT method achieves results on par with HA-FT in MCC metric, it falls short in robustness to input changes and is less precise than conventional HA-FT.

7 Discussion

In this section, we discuss the role of humans in classifier development. In this study, we showed that LLMs can effectively annotate datasets for a task as complex as lexical bias detection and that a downstream classifier achieves comparable results with a model trained on human-labeled data. However, the proposed framework only relies on a *near-unsupervised* regime. While annotations are automated by the LLMs, there are two crucial touchpoints of human interaction in the process. First, LLMs are selected based on their ranking on a dedicated human-labeled development set. Without this evaluation, practitioners will either have to rely on general NLP benchmarks, which may not reflect a good ranking for their specific task, or resort to random selection. Second, we prompt LLMs with human-labeled examples to enable in-context learning. Although this requires only a small number of annotations, it requires domain expertise and annotation effort. Lastly, the result of the behavioral testing shows a significant gap between the robustness of models trained on synthetic and human-made annotations. This indicates a need to improve the model's resilience to subtle changes in input. One possible way to tackle this is to augment the synthetic training process with human-made adversarial examples or increase the human effort in de-biasing the underlying dataset before the annotation (e.g., pruning/randomizing the named entities). While LLMs hold great potential, human intervention is still essential in automated annotation, especially for tasks such as lexical bias, both in the role of a guide and evaluator.

8 Conclusion

In this paper, we investigated the viability of using Large Language Models as annotators for training datasets to tackle the need for more high-quality resources in the lexical bias classification domain. We showed that general-purpose LLMs can generate reasonable annotations off-the-shelf, and we used three LLM annotators to create the first large-scale dataset for lexical bias classification - *Anno-lexical* - with 48330 sentences. We subsequently show that a classifier fine-tuned on the *Anno-lexical* synthetic annotations can match and even outperform a conventional model trained on human annotations while reducing the cost and effort required for human annotations. While our new model performs competitively on two lexical bias benchmarks, it falls short in classification precision and robustness against input perturbations. This defect becomes especially apparent when we scale down the size of *Anno-lexical* to match the size of the existing gold-standard dataset.

In our future work, we aim to evaluate the scaling laws of the synthetic annotations and the role of diversity in the underlying dataset. We hypothesize that the number of synthetic annotations can be exploited further, possibly leading to better and more robust models with the potential to transfer our results to other classification problems.

Limitations

As our approach to first annotate and then classify lexical bias relies directly on using state-of-the-art LLMs, one limitation is the computational cost of running the very large models. We acknowledge that our limited computational resources prevented us from testing and utilizing the most advanced models (those with more than 50 billion parameters and proprietary models). These cutting-edge models require immense computational power for inference but could potentially enhance performance. Secondly, we only evaluate the whole pipeline with three LLM annotators selected greedily based on the benchmark. We did not evaluate other combinations of the annotators due to the computational restrictions. However, since we evaluate the downstream model robustness and out-of-distribution generalization, another run with a random selection of LLMs would bring more insight into how the selection affects the downstream classifier behavior. While we present a generalizable pipeline and show that learning from synthetic data is viable for lexical bias detection—addressing a low-resource challenge in media bias research—the broader hypothesis of synthetic data utility extends beyond this domain. Our work provides a validated framework for resource-scarce tasks, but its verification on other tasks remains outside this paper's scope. We prioritize a proof of concept for media bias detection, offering both a scalable solution and inspiration for adapting this approach to other under-resourced domains.

Ethics Statement

Media bias strongly depends on personal perception, making it a sensitive issue, especially in the context of automated annotations. Some bias forms depend on factors other than the content, e.g., a different text perception due to a reader's background. While in this paper, we merely investigate the possibilities of automated data annotation if used within a publicly available classifier, quality control of what is classified as bias, especially when subjective, is a main part of our ongoing and future work. We recognize the potential for introducing bias in model training and annotation processes and have attempted to mitigate these through diverse data sources and balanced representation. We see no immediate risk to our work; however, we note that current models still make false predictions and discourage potential users from using them in production. By automating the annotation process, we aim to make the dataset creation in the lexical bias domain less expensive, which, together with additional quality control, will ideally lead to larger availability of media bias classifiers. We also believe that creating dedicated datasets and classifiers for individual tasks will result in lower energy consumption than running resource-expensive LLMs locally.

Lastly, we want to declare that the authors used ChatGPT during the writing process of this work, primarily for minor rephrasing and grammar correction.

Acknowledgements

This work was supported by the Lower Saxony Ministry of Science and Culture and the VW Foundation. Furthermore, this project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 554559555 and by EXIST-Gründungsstipendium. Finally, the authors would like to express their gratitude towards Prof. Dr. Michael Granitzer for consulting and valuable advices.

References

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-Source Large Language Models Outperform Crowd Workers and Approach ChatGPT in Text-Annotation Tasks. *Computation and Language*, (arXiv:2307.02179).

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Krahmer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood, Margot Mieskes, Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Chengliang Chai, Jiayi Wang, Nan Tang, Ye Yuan, Jiabin Liu, Yuhao Deng, and Guoren Wang. 2023. Efficient Coreset Selection with Cluster-based Methods. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pages 167–178, New York, NY, USA. Association for Computing Machinery.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6.

Michael Chmielewski and Sarah C. Kucker. 2020a. An mturk crisis? shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4):464–473.

Michael Chmielewski and Sarah C. Kucker. 2020b. An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. *Social Psychological and Personality Science*, 11(4):464–473.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In Plain Sight: Media Bias Through the Lens of Factual Reporting. 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 6342–6348, Hong Kong, China. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Laurence Gillick and Stephen J Cox. 1989. Some statistical issues in the comparison of speech recognition algorithms. In *International Conference on Acoustics, Speech, and Signal Processing,*, pages 532–535. IEEE.

Shahriar Golchin and Mihai Surdeanu. 2023. Time travel in llms: Tracing data contamination in large language models. *CoRR*, abs/2308.08493.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies (Volume 6: Industry Track), pages 165–190, Mexico City, Mexico. Association for Computational Linguistics.

Tomas Horych. 2022. Metody detekce vyvazenosti zpravodajskychch textu.

Tomas Horych, Martin Wessel, Jan Philip Wahle, Terry Ruas, Jerome Wassmuth, Andre Greiner-Petter, Akiko Aizawa, Bela Gipp, and Timo Spinde. 2024. Magpie: Multi-task analysis of media-bias generalization with pre-trained identification of expressions. In *"Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation"*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023*, pages 294–297, Austin TX USA. ACM.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2023. Analyzing Dataset Annotation Quality Management in the Wild.

David Krieger, Timo Spinde, Terry Ruas, Juhi Kulshrestha, and Bela Gipp. 2022. A Domain-adaptive Pre-training Approach for Language Bias Detection in News. In *2022 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Cologne, Germany.

Yuanyuan Lei and Ruihong Huang. 2024. Sentence-level Media Bias Analysis with Event Relation Graph.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Luyang Lin, Lingzhi Wang, Xiaoyan Zhao, Jing Li, and Kam-Fai Wong. 2024. IndiVec: An Exploration of Leveraging Large Language Models for Media Bias Detection with Fine-Grained Bias Indicators. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1038–1050, St. Julian's, Malta. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What Makes Good In-Context Examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active Learning Principles for In-Context Learning with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5011–5034, Singapore. Association for Computational Linguistics.

Catherine C. Marshall, Partha S.R. Goguladinne, Mudit Maheshwari, Apoorva Sathe, and Frank M. Shipman. 2023. Who Broke Amazon Mechanical Turk? An Analysis of Crowdsourcing Data Quality over Time. In *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci '23, pages 335–345, New York, NY, USA. Association for Computing Machinery.

MistralAI. 2024. Mixtral of experts: Dynamic gating for efficient model scaling.

Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

OpenAI. 2023. Gpt-3.5.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

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.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237:121641.

Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? disciplinary values in computer vision dataset development. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Timo Spinde, Smilla Hinterreiter, Fabian Haak, Terry Ruas, Helge Giese, Norman Meuschke, and Bela Gipp. 2023. The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias. *arXiv preprint arXiv:2312.16148*.

Timo Spinde, Christina Kreuter, Wolfgang Gaissmaier, Felix Hamborg, Bela Gipp, and Helge Giese. 2021a. Do You Think It's Biased? How To Ask For The Perception Of Media Bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 61–69.

Timo Spinde, David Krieger, Manu Plank, and Bela Gipp. 2021b. Towards A Reliable Ground-Truth For Biased Language Detection. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, Virtual Event.

Timo Spinde, Jan-David Krieger, Terry Ruas, Jelena Mitrović, Franz Götz-Hahn, Akiko Aizawa, and Bela Gipp. 2022. Exploiting transformer-based multitask learning for the detection of media bias in news articles. In *Proceedings of the iConference 2022*, Virtual event.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021c. Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Spinde, Lada Rudnitckaia, Sinha Kanishka, Felix Hamborg, Bela, Gipp, and Karsten Donnay. 2021d. MBIC – a media bias annotation dataset including annotator characteristics. In *Proceedings of the iConference 2021*, Beijing, China (Virtual Event).

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey.

Petter Törnberg. 2023. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning.

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. 2023. Llama 2: Open foundation and fine-tuned chat models.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Martin Wessel and Tomáš Horych. 2024. Beyond the surface: Spurious cues in automatic media bias detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 21–30, St. Julian's, Malta. Association for Computational Linguistics.

Martin Wessel, Tomás Horych, Terry Ruas, Akiko Aizawa, Bela Gipp, and Timo Spinde. 2023. Introducing MBIB - The First Media Bias Identification Benchmark Task and Dataset Collection. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2765–2774, Taipei Taiwan. ACM.

Endalkachew Yitayew. 2023. Flan-ul2 20b: Exploring google's latest language model.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Data Contamination

We tested the contamination of GPT-4-Turbo and GPT-3.5 Turbo by following (Golchin and Surdeanu, 2023) on a sample of the BABE dataset. Our tests indicated that no contamination is present for both models.

B Annomatic

Annomatic is a lightweight Python package designed to automate the process of text annotation using Language Models (LLMs). The package uses Haystack (https://github.com/deepset-ai/haystack) under the hood, and is suitable for integration into existing workflows or as standalone software.

Annomatic aims to implement two core principles: **model-agnosticism** - through enabling the user to use a variety of LLMs, and **task agnosticism** - to allow the user to define the annotation task freely. Currently, *Annomatic* features three model classes:

1. OpenAiModel
2. HuggingFaceModel
3. VllmModel

For integrating existing LLMs into the annotation workflow.

Annomatic is published under Apache-2.0 license and is available at https://github.com/Media-Bias-Group/annomatic. A snippet of a code using annomatic for data annotation is shown in Figure 4.

C Annolexical dataset

Table 6 (continues in Table 7) shows sentences sampled from the Annolexical dataset. First, we sample sentences related to the topic of Electric cars. We show two sentences - unbiased and unbiased for each of the five political leanings in the dataset (Left, Lean Left, Center, Lean Right, Right). Secondly, sentences on random topics are shown. Two examples - unbiased and bias per each one of the five most frequent news sources within the political leaning. The news sources are coded as shown Table 5. A brief statistics on the lengths of sentences in the dataset is shown at the end of the Table 7.

name	source	availability	parameters	link
Zephyr 7B beta	Tunstall et al. (2023)	open	7B	HuggingFaceH4/zephyr-7b-beta
OpenChat 3.5	Wang et al. (2023)	open	7B	openchat/openchat-3.5
Mistral-7B-v0.1 Instruct	Jiang et al. (2023)	open	7B	mistralai/Mixtral-8x7B-Instruct-v0.1
LLama 2 7B chat	Touvron et al. (2023)	open	7B	meta-llama/Llama-2-7b-chat
LLama 2 13B chat	Touvron et al. (2023)	open	13B	meta-llama/Llama-2-13b-chat
Falcon-7B-Instruct	Almazrouei et al. (2023)	open	7B	tiiuae/falcon-7b-instruct
Flan-T5-base	Chung et al. (2022)	open	248M	google/flan-t5-base
Flan-T5-large	Chung et al. (2022)	open	783M	google/flan-t5-large
Flan-T5-XL	Chung et al. (2022)	open	2.8B	google/flan-t5-xl
Flan-UL2	Yitayew (2023)	open	20B	google/ul2
Mixtral-8x7B Instruct	MistralAI (2024)	open	13B (MoE inference)	Mixtral-8x7B-Instruct-v0.1
GPT-3.5 Turbo	OpenAI (2023)	closed	unknown	openai/gpt-3-5-turbo
GPT-4 Turbo	OpenAI et al. (2024)	closed	unknown	openai/gpt-4-turbo

Table 4: All LLMs evaluated listed with meta data.

coding	link
altrnet	https://www.alternet.org/
vox	https://www.vox.com/
economist	https://www.economist.com/
conversation	https://theconversation.com/
cnbc	https://www.cnbc.com/
justthenews	https://justthenews.com/
bostonherald	https://www.bostonherald.com/
townhall	https://townhall.com/
nymag	https://nymag.com/
vogue	https://www.vogue.com/
theatlantic	https://www.theatlantic.com/world/
guardian	https://www.theguardian.com/
verge	https://www.theverge.com/
independent	https://www.independent.co.uk
wired	https://www.wired.com/
cnet	https://www.cnet.com/
csmonitor	https://www.csmonitor.com/
foreignaffairs	https://www.foreignaffairs.com/
nypost	https://nypost.com/
christanitytoday	https://www.christianitytoday.com/
christainpost	https://www.christianpost.com/
washingtontimes	https://www.washingtontimes.com/
hotair	https://hotair.com/
foxnews	https://www.foxnews.com/
newsmax	https://www.newsmax.com/
americanthinker	https://www.americanthinker.com/
newsbusters	https://www.newsbusters.org/

Table 5: Mapping of source company abbreviation in Table 6 and Table 7 to the original URL.

```python
import pandas as pd
from haystack.components.builders import PromptBuilder
from haystack.components.generators import HuggingFaceLocalGenerator

from annomatic.annotator import FileAnnotator

df = pd.read_csv("data_to_annotate.csv")

prompt = PromptBuilder(
    "Instruction: '{{text}}'\n"
    "Classify the sentence above as BIASED or NOT BIASED.\n"
    "Output: ",
)

# create annotator
model = HuggingFaceLocalGenerator(model="google/flan-t5-large")
annotator = FileAnnotator(
    model=model,
    output="./output.csv",
    batch_size=2,
    labels=["BIASED", "NOT BIASED"],
)

annotator.set_input(df, data_variable="text")
annotator.set_prompt(prompt)

# annotate the data and return as df
annotated_df = annotator.annotate(return_df=True)
```

Figure 4: Basic usage of the annomatic library for zero-shot annotation.

leaning	source	sentence	label
Electric cars			
L	alternet	Additionally, changes in consumer behaviour, such as a shift toward electric vehicles or renewable energy sources, could render certain infrastructure assets obsolete.	0
	vox	Auto unions and US carmakers recently smoothed over a huge pothole on the road to electric vehicles, but workers are facing much bigger ruts ahead on the route to clean energy.	1
LL	economist	EV s' low adoption rates, relative novelty and rapidly evolving technology, meanwhile, make it hard for buyers to tell how fast they lose their worth, which may put some off the purchase.	0
	conversation	Surely it won't appeal to the same core base of existing environmentally-conscious Tesla customers.	1
C	cnbc	CATL is a major supplier of electric car batteries for all the big industry players from BMW to Tesla.	0
	cnbc	Contemporary Amperex Technology , better known as CATL, is an understated giant in the electric vehicle world.	1
LR	justthenews	Michigan has fallen more than 1.9 million EVs short of reaching its climate goal of 2 million EVs driving on Michigan roads by 2030.	0
	bostonherald	Rarely do you get such a perfect visual of the shortcomings of the current electric vehicle market.	1
R	townhall	While it is true that 2030 is still over half a decade away, common sense dictates that a long-term trend of EV sales increases would be needed to reach this goal by that time.	0
	townhall	But not a single charger funded by the bipartisan infrastructure law has come online and odds are they will not be able to start powering Americans' vehicles until at least 2024.	1
Randomly sampled			
L	nymag	It's, at this point, a very long document, and was always the best part of writing the sketch.	0
		What happens when we sacrifice our humanity in the pursuit of a cohesive personal brand?	1
	vox	Phenylephrine isn't alone: Due to some quirks in the agency's history and policy, the FDA is engaged in an ongoing process of reviewing drugs that came on the market before its current efficacy standards went into effect.	0
		Fair Play is caustic and enthralling, but mostly it's the kind of movie that makes you wince with recognition - or, in any case, if you've ever made yourself small to avoid the rage of an insecure man.	1
	alternet	As a lobbyist, he worked his connection with Reid, who was then the No.	0
		It was chaotic and capricious at times, they said, but Trump gave second chances to many deserving inmates.	1
	vogue	For lips, Dries Van Noten's sculptural cases would make a chic addition to any makeup bag, as would Hailey Bieber's first makeup product-the editor-approved Rhode Peptide Lip Tint.	0
		Yeah, so our public media's been pretty good around this project, but interestingly, as Nikki said, there was a bit of hesitation at first from all of the media until news.com of all places ran the story in a very, what we saw a sympathetic way.	1
	theatlantic	This month, teens across the country have been adding their voices to the debate over in-person schooling, which has so far been dominated by adults-by parents, teachers, and politicians.	0
		Bannon was speaking with the former Trump-administration utility player Kash Patel, who promised to go after the free press.	1
LL	economist	In Sweden, where corporate debt is an eye-watering 155% of GDP , the average effective interest rate on outstanding bank loans to companies rose from 1.5% to 3.9% in the year to March and will be higher still by now.	0
		It is hard, for instance, to see Mr Putin in the same way after hearing of his childlike obsession with ice-cream.	1
	conversation	We would simply need to ditch the pension income test in cases where that income came from paid work, leaving the assets test in place.	0
		Bettmann via Getty Images August 1, 2023 Why American culture fixates on the tragic image of J. Robert Oppenheimer, the most famous man behind the atomic bomb Complex as they are, Oppenheimer's life and views of the bomb are far easier to wrestle with than the reality of nuclear power itself.	1
	guardian	If there is a big difference between someone critiquing your scrummaging or decisions in the morning paper and receiving abuse and death threats from faceless accounts worldwide, there are clearly issues for media companies as well.	0
		It's not going to work with cold, grumpy Britons, is it?	1
	verge	Governments, philanthropies, and the private sector unveiled over $1 billion dollars in new catalytic grant funding for methane reduction since COP27.	0
		The Greatest Killer in New Orleans Wasn't the Hurricane. It Was the Heat.	1
	independent	Las Vegas is set to host its first grand prix for more than 40 years (AP) From McLaren partners Hilton to Red Bull sponsor PokerStars everyone wants in on the Las Vegas F1 action, tying their businesses to the glamour sport's most glamorous event through local activations and a massive global television audience.	0
		From glorious afternoon teas to exquisite dinner dishes (including a six-course tasting menu), this is a fabulous foodie destination, too.	1

Table 6: This table shows random samples from the anno-lexical dataset. First, samples concerning topic of Electric cars are shown, two examples - unbiased and biased - are shown for each of the political leaning. Secondly, sentences on random topic are shown. Two examples - unbiased and bias per eahc of the five most frequent news sources within the political leaning. Political leaning is coded as: L - Left, LL - Lean Left, C - Center, LR - Lean Right, R - Right. The labels are coded as 0 - neutral, 1 - lexical bias. The table continues on a next page.

leaning	source	sentence	label
C	wired	We're going to be talking about the nature of time, the most familiar and the most mysterious quality of the physical universe.	0
		Skin all over the body starts to sag, underneath your arms, breast issues, your scrotum starts to sag.	1
	cnet	Risks of vitamins for eyes Most vitamins and supplements are generally considered safe for people to take, as they're nutrients that your body naturally requires.	0
		Navigating the highly competitive solar industry requires avoiding the less-than-honest companies out there.	1
	csmonitor	The government accused the firm of turning a blind eye to transactions made by Hamas, Palestinian Islamic Jihad, Al Qaeda, and ISIS.	0
		And this shifts the generation's choices in everything from what they eat to where they work.	1
	foreignaffairs	With Saudi and Jordanian blessing, Qatar, the UAE, and Egypt have agreed on a scenario in which figures such as Ismail Haniyeh, Hamas's Doha-based political leader, would play a role in a Palestinian coalition government proposed at the joint summit between the Arab League and the OIC.	0
		Putin has already made one historic miscalculation; no one can be sure he will not make another.	1
	cnbc	Discovery, he discussed Paramount Global as an example of a company whose prospects seem shaky.	0
		Not to mention, the studio inundated Disney+ with series in an effort to pad its platform, making some fans feel like they had to slog through hours of stories in order to understand what was happening in the films.	1
LR	nypost	QVC When we first reviewed the Bissell SpinWave Cordless Hard Floor Spin Mop, our jaws nearly dropped to the floor.	0
		We don't think Ginger is even trying to gain back our trust.	1
	christianitytoday	Oh man, I've seen some weird and awkward stuff, some of them are probably not appropriate.	0
		But if you were to, as the pastor, get up in the pulpit and say, Stop going to the mountains for camping, why in the world would you do that.	1
	christianpost	The situation in North Carolina is not the first example of a parent facing consequences after a weapon ended up in a minor's hands.	0
		We should think on those words often and ask if we know the terror of the Lord, and if we are then persuading men.	1
	washingtontimes	NASHVILLE, Tenn. (AP) - Major League teams stocked up on pitching Wednesday in the big league phase of the winter meeting draft of unprotected players, taking pitchers with eight of the 10 selections - including three from the New York Yankees' system.	0
		Supporters say Mr. Biden has been unfathomably unlucky, buffeted by global events that would have challenged anyone in office.	1
	hotair	Everyone was heartsick, and no one could talk about anything but the tragedy on base.	0
		There were public spats about ending annual passes for Reedy Creek employees (subsequently replaced by bonuses more than equal to the price of a pass), squabbles about trimming $8 million from Disney's silky smooth the roads-paving budget, and a report last week that - citing this perceived cronyism - morale is sinking fast among Reedy Creek employees who stayed through the transition.	1
R	foxnews	EX-NFL PLAYER SERGIO BROWN PLEADS NOT GUILTY TO KILLING HIS MOTHER Past media guides say Patel was a financial planning and analysis coordinator and then manager.	0
		All these communications were conducted with an alias email account, something you're not supposed to do in order to prevent future discovery in like a court of law.	1
	newsmax	Rothstein, the Johns Hopkins researcher, plans to prescribe the drug to his patients.	0
		Strict scrutiny, to borrow a phrase, is what Israel always faces.	1
	townhall	Earlier this year, a Massachusetts couple had their foster care application denied because they did not sufficiently support transgender ideology.	0
		Of course, one can look at the news, see the corruption and the internal decay in some of our institutions and reasonably understand why individual Americans may not have much interest in serving a higher purpose.	1
	americanthinker	Except for 2016, this is the highest level reported by the public since Gallup initiated this time series in 2000.	0
		Our children are being assailed in classrooms if they admit to being a male or a female, a distinction fundamental to the human psyche regarding identity - not to mention the invasion of morality that occurs through the media every day, visibly, in front of the same vulnerable population.	1
	newsbusters	A driver behind that could be the median age which is just 30.	0
		On its website, Absolut confirmed that it was the first brand to advertise in LGBTQ+ magazines back in 1981 and since then has stood firm in its support for the gays.	1
Statistic			
	# sources: 118	Average sentence length (words): 23.7 ± 10.69	

Table 7

You are an expert in media bias.
{ for TEXT, LABEL, EXPLANATION in examples }
Instruction: '[TEXT]'
Classify the sentence above as BIASED or NOT BIASED.
Output: Let's think step by step. [EXPLANATION] The answer is [LABEL].
{ endfor }
Instruction: '[SENTENCE]'
Classify the sentence above as BIASED or NOT BIASED.
Output: Let's think step by step.

Table 8: Prompting Template in pseudo-code. {..} indicates a command.

Citation for this Paper

```
@inproceedings{Horych2025,
 author={Tomas Horych, Christoph Mandl, Terry Ruas, Andre Greiner-Petter, Bela Gipp, Akiko
Aizawa, Timo Spinde},
 title={The Promises and Pitfalls of LLM Annotations in Dataset Labeling: a Case Study on
Media Bias Detection},
 booktitle={Findings of the 2025 Conference of the The Nations of the Americas Chapter of
the Association for Computational Linguistics: NAACL 2025},
 pages={17},
 publisher={Association for Computational Linguistics},
 address={Albuquerque, USA},
 year={2025},
 month={01}
}
```