Biases in Large Language Model-Elicited Text: A Case Study in Natural Language Inference

Grace Proebsting
Haverford College
gproebstin@haverford.edu

Adam Poliak
Bryn Mawr College
apoliak@brynmawr.edu

Abstract

We test whether NLP datasets created with Large Language Models (LLMs) contain annotation artifacts and social biases like NLP datasets elicited from crowd-source workers. We recreate a portion of the Stanford Natural Language Inference corpus using GPT-4, Llama-2 70b for Chat, and Mistral 7b Instruct. We train hypothesis-only classifiers to determine whether LLM-elicited NLI datasets contain annotation artifacts. Next, we use point-wise mutual information to identify the words in each dataset that are associated with gender, race, and age-related terms. On our LLM-generated NLI datasets, fine-tuned BERT hypothesis-only classifiers achieve between 86-96% accuracy. Our analyses further characterize the annotation artifacts and stereotypical biases in LLM-generated datasets.

1 Introduction

Creating NLP datasets with Large Language Models (LLMs) is an attractive alternative to relying on crowd-source workers (Ziems et al., 2024). Compared to crowd-source workers, LLMs are inexpensive, fast, and always available. Although LLMs require validation (Pangakis et al., 2023), they are an efficient tool to annotate data (Zhao et al., 2022; Bansal and Sharma, 2023; Gilardi et al., 2023; He et al., 2024). In addition to relying on LLMs for data annotation, researchers can elicit text from LLMs to create NLP datasets. For instance, LLMs have been used to generate training sets for NLP classification tasks like sentiment and intent classification (Ye et al., 2022; Sahu et al., 2022; Chung et al., 2023; Møller et al., 2024).

Eliciting text from humans can yield NLP datasets with stereotypical biases (Rudinger et al., 2017) and annotation artifacts (Cai et al., 2017; Kaushik and Lipton, 2018). Since researchers use LLMs to create textual datasets, we study whether LLM-elicited datasets similarly suffer from stereotypical biases and annotation artifacts. To compare human- and machine-elicited textual data, we create LLM-generated versions of the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015) by providing LLMs with the same instructions given to SNLI crowd-source workers.

We focus on Natural Language Inference (NLI), the task of determining whether a hypothesis sentence could be likely inferred from a premise (Dagan et al., 2005), since popular NLI datasets with crowd-sourced hypotheses contain biases. We apply standard approaches to detect annotation artifacts in NLI by training hypothesis-only classifiers and identifying words highly associated with specific NLI labels. Further, we search for race, age, and gender-based stereotypical biases by finding words most associated with these social groups, and compare them with biases in SNLI.

We find that LLM-elicited NLI contains both hypothesis-only and social biases. On our LLM-generated NLI datasets, fine-tuned BERT classifiers achieve 86-96% accuracy when given only the hypotheses, compared to 72% performance on SNLI. We also find the LLM-generated datasets contain similar gender stereotypes as SNLI. Our research suggests that while eliciting text from LLMs to generate NLP datasets is enticing and promising, thorough quality control is necessary.

2 Background & Motivation

There is a robust literature focusing on whether LLMs contain biases (Nozza et al., 2021; Sheng et al., 2021; Mei et al., 2023; Kolisko and Anderson, 2023; Gallegos et al., 2024; Liu et al., 2024; Shin et al., 2024; Raj et al., 2024; Hu et al., 2024). We similarly evaluate biases in LLMs, but our focus is different: specifically, we ask whether LLMs are a suitable replacement for crowdsource workers when creating NLP datasets. Concretely, we investigate whether NLP datasets with LLM-elicited

Proceedings of the 31st International Conference on Computational Linguistics, pages 5836–5851
January 19–24, 2025. ©2025 Association for Computational Linguistics

Premise	Two women are hiking in the wilderness.	
	Entailment	**Contradiction**
SNLI	There are two women outdoors.	There are two women in the living room.
Llama	There are people outdoors.	A couple is having a picnic in a park.
Mistral	There are people in nature.	The women are shopping for clothes.
GPT-4	People are outdoors.	Two women are swimming in a pool.

Table 1: Entailed and contradicted hypotheses produced by humans (SNLI) and three LLMs (Llama-2 70b for Chat, Mistral 7b Instruct, and GPT-4) in response to the same premise.

text contain similar annotation artifacts and social biases as NLP datasets with human-elicited text.

Prompting humans to generate text for large-scale NLP datasets can lead to biased datasets. Famously, datasets for the Story Cloze Test and NLI contain biases introduced by their human elicitation protocols. To create a dataset for the Story Cloze Test, i.e. the task of determining the correct ending of a story, Mostafazadeh et al. (2016) asked crowd-source workers "to write novel five-sentence stories." Bowman et al. (2015) created SNLI by providing crowd-source workers image captions from the Flickr30k corpus (Young et al., 2014) and instructing workers to write three alternative captions: one that is *definitely true*, one that *might be true*, and one that is *definitely false*. These human-elicitation protocols are responsible for creating 1) annotation artifacts that enable naive models ignoring substantial context to perform surprisingly well (Schwartz et al., 2017; Tsuchiya, 2018; Gururangan et al., 2018; Poliak et al., 2018; Feng et al., 2019), and 2) social biases that "amplify . . . stereotypical associations" (Rudinger et al., 2017).

In addition to these concerns, creating datasets by eliciting text from humans can be expensive. LLMs can efficiently generate, label, and clean datasets for a wide variety of applications (Ziems et al., 2024). LLMs have been used to generate instruction-tuning datasets (Honovich et al., 2023; Wang et al., 2023; Peng et al., 2023), synthetic versions of benchmarks like SuperGLUE (Wang et al., 2019; Gupta et al., 2024), counterfactuals for dataset augmentation (Wu et al., 2021; Chen et al., 2023), attributable information seeking (Kamalloo et al., 2023), and free-text classification explanations (Wiegreffe et al., 2022). LLM-elicitation is especially attractive for sensitive domains, e.g. clinical NLP, where datasets must not leak personal identifying information (Frei and Kramer, 2023; Xu et al., 2024b). LLMs-elicited text is pervasive even among crowd-source workers: Veselovsky et al. (2023) claim that "33–46%" of the crowd-source workers hired for a summarization task likely used LLMs to produce summaries.

Some LLM-generated datasets involve no post-filtering step (Peng et al., 2023; Xu et al., 2024a,b). However, most resources built with LLM-elicitation include thorough quality assurance, either through "human-in-the-loop" curation (Wiegreffe et al., 2022; Liu et al., 2022; Kamalloo et al., 2023), statistical filtering (Wu et al., 2021; Ye et al., 2022; Wang et al., 2023) or relying on neural models to filter LLM-generated data (Wiegreffe et al., 2022; Chen et al., 2023; Yehudai et al., 2024; Gupta et al., 2024). While we advocate for filtering steps to ensure quality and remove biases in LLM-elicited text, we focus on analyzing the unfiltered output of "out-of-the-box" LLMs for NLP datasets. We ask, specifically in the context of NLI, whether LLM-elicited text contains biases, and if so, what are these biases?

3 Creating LLM-Elicited NLI

We use NLI as a case study to explore whether LLM-generated text contain similar biases as human-written text since human-elicited NLI datasets contain annotation artifacts and stereotypical social biases. We create modified versions of SNLI by prompting LLMs with the same instructions that Bowman et al. (2015) gave to crowd-source workers. Table 1 provides examples from each dataset. We further verify the quality of the generated hypotheses and determine how different they are from those in SNLI.

LLMs under consideration We select a diverse set of LLMs for dataset generation: **GPT-4** (OpenAI, 2023), **Llama-2 70b for Chat** (Touvron et al., 2023), **Mistral 7b Instruct** (Jiang et al., 2023), and **PaLM 2 for Chat** (Anil et al., 2023). [1] These mod-

[1]For GPT-4 we use `gpt-4-0613`, for Llama Chat 70b we use `llama-2-70b-chat`, for Mistral 7b Instruct we

Data set sizes:	
Training pairs	133,629
Evaluation pairs	6,525
Hypothesis mean token count:	
SNLI train	8.1
Llama train	9.4
Mistral train	9.1
GPT-4 train	9.2
PaLM 2 train	7.7
Mean Jaccard similarity with SNLI:	
Llama train	0.19
Mistral train	0.22
GPT-4 train	0.20
PaLM 2 train	0.25

Table 2: Summary statistics for each dataset.

	Overall	Entail	Neutral	Contra
SNLI	92.7	87.0	95.0	96.0
Llama	89.7	73.0	98.0	98.0
Mistral	83.7	70.0	91.0	90.0
GPT-4	94.3	84.0	99.0	100.0
PaLM 2	77.0	62.0	90.0	79.0

Table 3: Percentage of examples where we agreed with the label of 300 NLI example pairs from each dataset.

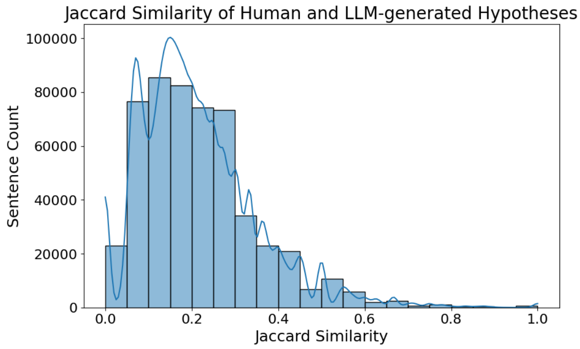


Figure 1: Frequency (y-axis) of lexical overlap (x-axis) between LLM and corresponding SNLI hypotheses.

els vary in parameter count, parent company, and training technique. We initially included models with open training sets to test for data contamination, e.g. AI2's OLMo-7B-Instruct (Groeneveld et al., 2024), DataBrick's dolly-v2-12b (Conover et al., 2023) or EleutherAI's gpt-j-6b (Wang and Komatsuzaki, 2021), but these open-data models did not create accurate entailed hypotheses in initial experiments. Given computational constraints, we were unable to use LLMs, e.g. BLOOM (Workshop et al., 2022) or Falcon (Almazrouei et al., 2023).

Dataset generation To mirror Bowman et al. (2015)'s dataset elicitation pipeline, we prompted LLMs with the same instructions provided to crowd-source workers for SNLI.[2] To balance lexical diversity with reproducibility, we set the temperature and top-p respectively to 0.75 and 0.9 for all LLMs. Additionally, we use the default top-k parameter for each LLM. Due to budget constraints, for each LLM, we create hypotheses for a third of the premises in the SNLI train set and all premises in the SNLI evaluation set. Table 2 contains statistics regarding each dataset.

Dataset validation To verify the LLMs correctly generated hypotheses for each label, we sampled 100 premises and manually verified the labels for the corresponding 300 NLI sentence pairs for each model. Table 3 reports our agreement with the NLI labels for each LLM. Since we agreed with less than 80% of the examples sampled from the PaLM2-elicited dataset, we do not consider the dataset generated by PaLM2 in our later studies.

To ensure the LLM-generated hypotheses are not simply memorized and copied verbatim from SNLI, we compute the Jaccard similarity of the words within pairs of LLM-generated and SNLI hypotheses corresponding to the same premises and labels.[3] Figure 1 plots the distribution of the Jaccard similarities between SNLI and corresponding LLM-generated hypotheses. Table 2 reports the average Jaccard similarity for each *individual* LLM dataset. **LLM and human-generated hypotheses have low lexical overlap**, demonstrating that these LLMs do not copy SNLI verbatim.[4]

4 Study 1: Hypothesis-Only Artifacts

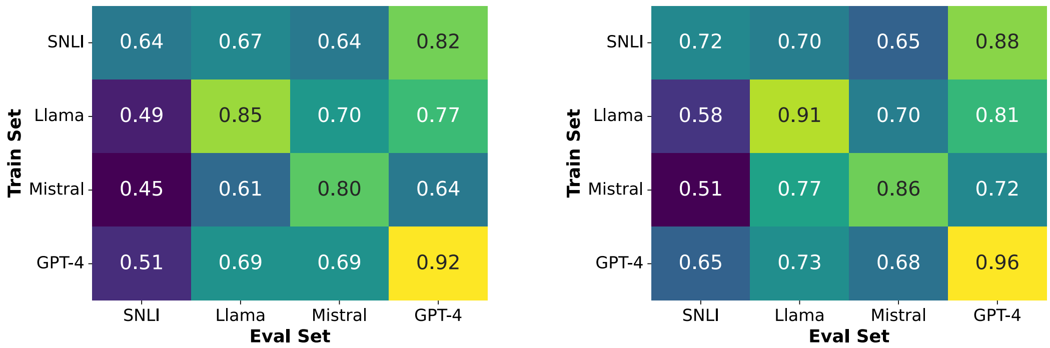
In our first study, we determine whether LLM-elicited NLI datasets contain annotation artifacts

use `mistral-7b-instruct-v0.1`, and for PaLM 2 for Chat we use `chat-bison`.

[2]We slightly changed the prompt to ensure the LLM's output was valid JSON. We provide the full prompt in the Appendix (Figure 6).

[3]Jaccard similarity is a measure of set overlap that ranges between 0.0 (a disjoint set) and 1.0 (an identical set).

[4]Reviewers noted the limits of Jaccard similarity since LLMs might paraphrase hypotheses from SNLI if the LLMs were pre-trained on SNLI. A manual review of thousands of examples suggested that these LLM-generated hypotheses contained semantically different content from that of the hypotheses in SNLI, i.e., the LLM-generated hypotheses were not merely paraphrased from SNLI.



(a) Naive Bayes

(b) BERT-based

Figure 2: Accuracy of each hypothesis-only classifier on each LLM and human-generated evaluation set. Each row represents the hypothesis-only NLI dataset used for training, and each column represents the evaluation dataset.

that allow hypothesis-only models to outperform a majority-class baseline. We train two types of hypothesis-only models: Naive Bayes (NB) using the case-sensitive implementation from scikit-learn with unigram features (Pedregosa et al., 2011), and a fine-tuned BERT classifier (Devlin et al., 2019), specifically bert-base-uncased models with 3-class sequence classification heads and default Hugging-Face hyper-parameters (Wolf et al., 2020),[5] which we train for 1 epoch using AdamW (Loshchilov and Hutter, 2018), a learning rate of 2e-5, a weight decay of 0.01, and a batch size of 16.

We train hypothesis-only models on each of our train sets (3 LLM-generated and the filtered SNLI) and evaluate them on all evaluation sets. Figure 2 reports the accuracy of the hypothesis-only models.

The highest-performing model on each evaluation set was trained on the corresponding train set - in each column in Figure 2, the highest accuracy is along the diagonal. Surprisingly, the SNLI-trained models perform much better on the GPT-4 generated evaluation set (0.82 for NB and 0.88 for BERT) than on the SNLI evaluation set (0.64 for NB and 0.72 for BERT), indicating that GPT-4 might contain similar annotation artifacts as SNLI.

We also notice that hypothesis-only models trained on LLM-generated data perform much better on other LLM-elicited datasets than on SNLI, as the accuracies in the first column are much lower than the other columns in both figures. This might indicate that the LLMs produce similar biases.

Qualitative analysis of give-away words The NB models with unigram features significantly outperform a majority baseline (Figure 2a), indicating that the hypotheses contain *give-away words*—single words that are highly indicative of a label.

We identify give-away words for each train set by calculating the conditional probability of each label l given the presence of a word w in a hypothesis: $p(l|w) = \frac{count(w,l)}{count(w)}$. We consider all give-away words with a conditional probability of at least 0.8. We follow Poliak et al. (2018) and sort give-away words by their frequency "since this statistic is perhaps more indicative of a word w's effect on overall performance compared to $p(l|w)$ alone." Table 4 reports the top 10 give-away words for each label in all train sets.

Entailed examples in SNLI often contain generic words like *humans*, *activity*, and *interacting*. We find a similar pattern in LLM-generated entailed hypotheses, e.g. *person* and *activity* in GPT-4 and Llama. Unlike in SNLI, the capitalized word *There* is a give-away for LLM-elicited entailed examples. LLMs often copy features from examples in prompts (Elhage et al., 2021; Olsson et al., 2022; Bansal et al., 2023; Zhang et al., 2024), which might explain why *There* is a give-away word in these LLM-elicited datasets. Human-generated neutral hypotheses often contain modifiers (*tall, sad, professional*) and superlatives (*first, favorite, winning*). LLMs similarly add embellishing details about emotions or intentions (*enjoying, fun, practicing, trying*) or the relationships between agents (*friends, couple, team*) that are not explicit in the premise. Two of Llama's neutral give-away words, *Someone* and *catch*, appear in the prompt's example of a neutral hypothesis.

Lastly, both human- and LLM-elicited contra-

[5] We did not perform hyper-parameter tuning since our goal is simply to establish whether a hypothesis-only model can perform well on an LLM-elicited NLI dataset.

	Word	$p(l\|w)$	Freq	Word	$p(l\|w)$	Freq	Word	$p(l\|w)$	Freq
SNLI	Humans	0.95	128	tall	0.85	418	sleeping	0.84	1747
	least	0.92	78	sad	0.81	322	Nobody	0.93	592
	activity	0.83	47	first	0.87	298	asleep	0.83	523
	multiple	0.81	37	owner	0.83	284	couch	0.81	477
	interacting	0.85	34	birthday	0.83	227	naked	0.88	248
	motion	0.97	32	winning	0.88	186	tv	0.81	207
	physical	0.83	30	favorite	0.88	180	cats	0.89	199
	occupied	0.8	15	professional	0.83	149	TV	0.81	177
	balances	0.82	11	vacation	0.94	141	No	0.93	134
	consuming	0.8	10	win	0.86	140	television	0.83	124
Llama	person	0.81	22264	Someone	1	4092	celebrity	0.92	2359
	People	0.86	7059	trying	0.9	3023	actually	0.94	2075
	standing	0.84	4359	going	0.95	1604	cat	0.9	1973
	outdoors	0.93	2390	break	0.87	1339	Everyone	0.93	1913
	engaging	0.94	1689	fun	0.88	1165	adult	0.89	1782
	Three	0.92	1593	practicing	0.86	1142	fashion	0.85	1766
	gathered	0.93	1513	ride	0.82	811	red	0.84	1537
	activity	0.83	1412	or	0.83	795	signing	0.92	1437
	public	0.82	1230	discussing	0.88	720	autographs	0.93	1398
	vehicle	0.87	1185	catch	0.95	622	sleeping	0.82	1371
Mistral	There	0.99	16707	be	0.97	5154	The	0.81	38491
	outdoors	0.87	1055	trying	0.8	4875	sitting	0.83	14564
	three	0.83	720	may	0.98	3815	bench	0.87	8545
	four	0.88	335	having	0.85	2039	not	0.94	8068
	urban	0.83	318	going	0.83	1877	subject	0.87	3672
	consuming	0.94	217	or	0.86	1858	couch	0.91	2330
	multiple	0.83	211	friends	0.95	1499	empty	0.89	1433
	vertical	0.84	182	It	0.9	1486	cards	0.92	1171
	acrobatic	0.88	176	could	0.98	1311	no	0.92	955
	many	0.87	153	fun	0.92	1201	movie	0.9	938
GPT-4	person	0.85	11764	to	0.85	7087	swimming	0.92	16281
	outdoors	0.97	8182	for	0.89	5791	pool	0.91	14638
	individual	0.96	4569	his	0.82	5042	reading	0.8	3492
	individuals	0.89	3878	friends	0.94	3439	book	0.81	3048
	There	0.86	3794	enjoying	0.85	2073	sleeping	0.91	2326
	Individuals	0.97	2159	couple	0.81	1878	cooking	0.84	2126
	interacting	0.98	1377	from	0.82	1823	cat	0.9	1875
	activity	0.97	1250	taking	0.82	1093	dress	0.8	1537
	gathered	0.88	1248	practicing	0.87	1092	alone	0.94	1293
	public	0.85	976	team	0.88	972	library	0.91	1274

(a) entailment (b) neutral (c) contradiction

Table 4: The most highly correlated words for each train set for given labels (the columns (c), (d), and (e)), thresholded to those with $p(l|w) >= 0.8$ and ranked according to frequency.

dicting hypotheses contain negation words, e.g. *nobody*, *no*, *not*. As noted by Poliak et al. (2018), premises "sourced from Flickr naturally deal with activities." Therefore, similar to how contradicted hypotheses in SNLI often mention *sleeping*, it is not surprising that LLM-elicited contradictions mention actions that cannot occur simultaneously to the action in the premise, e.g. *swimming* for GPT-4 and *sitting* for Mistral. Further, these verbs often occur in frequently repeated phrases that negate an

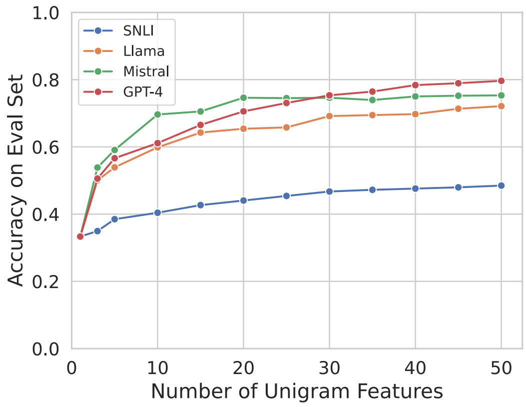


Figure 3: Accuracy of NB models using only the n "most informative" unigram features for each train set evaluated on its corresponding evaluation set.

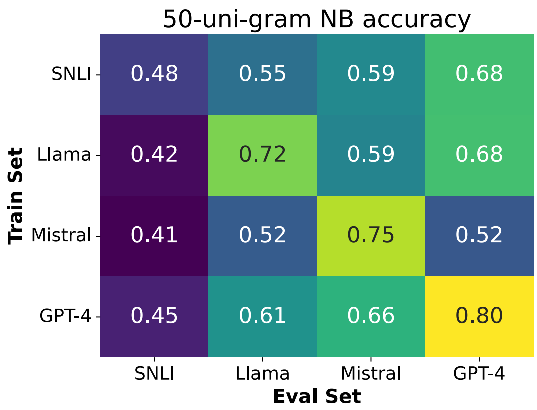


Figure 4: Accuracy of NB models with only the fifty most informative unigram features from their train set.

action described in the premise. For example, the phrases *"swimming in a pool"* and *"sitting on a bench"* respectively occur more than 10,000 times in the GPT-4 and Mistral-generated train sets.

Few unigrams needed for high NB accuracy. How many give-away words are necessary to accurately classify LLM-elicited NLI? To study this question, we train NB models that *only receive the **n** most informative give-away words as features.* We find the most informative words for each train set by performing a chi-squared test on all words with respect to each label. We threshold to the top n most informative unigrams and use only these words to train each n-feature NB model.

Figure 3 reports the accuracy of NB hypothesis-only models using just 1 to 50 features. Compared to SNLI, the LLM-elicited datasets are far easier to classify using a sparse selection of unigram features. For example, with just 10 unigrams, all LLM-trained NB models achieve greater than 60% accuracy, while the SNLI-trained 10-feature NB model only narrowly outperforms the majority-class baseline. This result indicates that LLM-generated hypotheses are trivial to classify not only due to the simplicity of the necessary features (unigrams) but also because only a negligibly small number of these simple features are required.

Figure 4 reports the accuracy of 50-unigram–feature NB models when evaluated on all four evaluation sets. NB models trained with sparse unigram feature sets on the LLM-generated hypotheses outperform a random baseline on the evaluation sets of the other LLM-generated hypotheses. This suggests that highly informative unigram features from one LLM-elicited dataset can be informative on the other LLM-elicited datasets. Additionally, like the NB and BERT-based hypothesis-only models trained on the entire feature set, the 50-feature NB hypothesis-only model trained on SNLI performs better on the GPT-4 evaluation than the SNLI evaluation set. Overall, these results suggest that the high accuracy of full-feature NB models across the evaluation sets might be attributed to a sparse set of give-away words that are common across the LLM-elicited datasets.

5 Study 2: Stereotypical Biases

Our second study analyzes whether LLM-elicited versions of SNLI, like the human-elicited SNLI, contain stereotypical social biases. Following Rudinger et al. (2017), we use pointwise mutual information (PMI) to identify words in each dataset that are most associated with gendered, racial, or age-based terms. Given word w_1 and w_2, the PMI between w_1 and w_2 is $\log(\frac{p(w_1,w_2)}{p(w_1)p(w_2)})$. For each dataset, we find the top co-occurring words in hypotheses by PMI with race, gender, and age-related query words that co-occur at least 3 times.

Gender-based stereotypes. Table 5 reports the top PMI terms for *man, men, woman* and *women.* PMI results for all query words can be found in the Appendix. In both the human-elicited and LLM-elicited datasets, male query words are associated with violence, work, and physical activity. In SNLI these terms include *burns, surfs, compete, wrestling, suits, poker, uniforms, chess, cars*. In the LLM-elicited datasets, terms highly associated with male terms include *suit, mowing, basketball, golf, cutting, boxing, sparring,* and *fighting*.

SNLI

woman mascara† knits‡ applies‡ sleds lipstick makeup‡ secret knitting‡ scarf countryside
man burns surfs nose buys container internet orders tractor popcorn dives
women cakes‡ yoga praying volleyball dresses thinking fruit tea talking‡ dance
men burn compete† wrestling suits† poker celebrate passing uniforms chess cars

GPT-4

woman ballgown gala bikini‡ oven‡ dress‡ ballroom‡ cookies‡ baking‡ heels‡ button
man spiderman shaving‡ suit‡ mowing‡ hamburger tuxedo beard‡ tie‡ proposing frowning‡
women dresses‡ mall‡ yoga† tea shopping‡ relaxing† picnic‡ sunbathing‡ baking dancing†
men suits‡ hats laying installing football‡ hard basketball‡ gym‡ rodeo skyscraper‡

Llama

woman lap‡ makeup‡ applying‡ arms‡ nails‡ mirror‡ sink knitting‡ sunbathing‡ flower†
man shaving‡ basketball‡ beard‡ guitar‡ girlfriend three golf‡ stadium‡ walks‡ ironing
women tea† clothing‡ socializing smiling each other routine party standing dancing
men football‡ dark field‡ basketball‡ instruments games‡ video‡ inside room playing‡

Mistral

woman cradling‡ arms sewing baby‡ flower† newborn serving gymnastics herself her‡
man diving† thrown net western tame horse wild his‡ cutting swinging‡
women japanese† traditional† clothes talking groceries posing conversation shopping smiling relaxing
men boxing‡ suits‡ robes ring sparring† fighting‡ court football basketball‡ match

Table 5: Top-ten words in hypothesis by PMI with gender-related query words in the same hypothesis, filtered to co-occurrences of at least three. (Hypothesis words that also appear in the premise are not included.) Significance of a likelihood ratio test for independence denoted by † ($\alpha = 0.01$) and ‡ ($\alpha = 0.001$).

In SNLI, the female query words are associated with physical appearance (*mascara, lipstick, makeup, dresses*) and leisure activities (*knits, yoga, cakes, tea, talking, dance*). LLM-generated hypotheses display similar stereotypes: female query words are related to domesticity (*oven, cookies, baking, knitting, cradling, baby, sewing, groceries*) and leisure activities (*mall, yoga, tea, shopping, relaxing, picnic, sunbathing, dancing, socializing, party, talking*). In the LLM-elicited datasets, female query words are also associated with clothing and physical appearance (*bikini, dress, heels, lap, makeup, arms, nails, clothing*).

Label-specific gender biases. To study how stereotypical biases appear based on NLI labels, for each NLI label, we now compute the PMI of hypothesis words with query words that appear in the premise. This allows us to determine if the LLMs contain stereotypical biases that are specific to different NLI labels. Table 6 reports label-specific biases for gender-related queries.

Broadly, LLM-generated entailed and neutral hypotheses display similar biases as the overall PMI results: male query words are associated with violence, physicality, and work, e.g. *workers, military, soldiers*, while female query words are associated with leisurely or domestic activities and physical appearance, e.g. *quilt, party, beauty*. A notable exception is that both Llama and Mistral associate "woman" with *scientist* and GPT-4 associates "woman" with *businesswoman*.[6] Additionally, Llama and Mistral associate "women" with *sporting* and *athletes*, respectively.[7]

Both human and LLM-generated *contradictions* sometimes flip the gender of the subject between the premise and hypothesis. In SNLI contradictions, male premise words are associated with *ladies* and *wife*, and LLM-generated contradictions feature *bikini* and *women*. Similarly, female

[6] Respectively entailment and neutral columns in Table 6.
[7] Entailment column in Table 6.

Query	ENTAILMENT	NEUTRAL	CONTRADICTION
man	**SNLI**: often gun climbs a‡ seated	**SNLI**: stops bald cowboy cafe newspaper	**SNLI**: gas scooter wife sings wears
	GPT-4: bathroom firearm casual embracing machine	**GPT-4**: latte cigar warehouse guy‡ adventurer	**GPT-4**: café bikini hat dolphins formal
	Llama: entertaining his‡ paper wood father‡	**Llama**: article summit fan avoid seafood	**Llama**: waters packed negotiating kidnapping before
	Mistral: presentation romantic moment a‡ scaling	**Mistral**: conference debris board summit a‡	**Mistral**: shirt costume tie a‡ individual‡
men	**SNLI**: workers guys† ball several they	**SNLI**: businessmen† crew workers† charity construction†	**SNLI**: ladies break party enjoying lunch
	GPT-4: workers construction‡ machinery project site	**GPT-4**: guys‡ foundation industrial soldiers‡ workers‡	**GPT-4**: individuals‡ playground‡ women‡ people‡ everyone‡
	Llama: parade† marching‡ industrial formal construction‡	**Llama**: cowboys‡ soldiers‡ complex† fishermen workers‡	**Llama**: awards† ballet celebrities‡ players‡ parade
	Mistral: fishermen‡ workers‡ job† military† personnel	**Mistral**: workers‡ soldiers‡ cowboys long-distance vendors	**Mistral**: casual admiring dressed‡ they‡ already
woman	**SNLI**: her‡ touching lady a‡ women	**SNLI**: herself husband† dress won clothes	**SNLI**: feeding a‡ phone she nothing
	GPT-4: female‡ stand exiting lady‡ toys	**GPT-4**: quilt‡ businesswoman bag lady‡ casual	**GPT-4**: lady‡ suit‡ man‡ a‡ dinner
	Llama: scientist mother‡ her customer off	**Llama**: savoring meditating furry considering hiker	**Llama**: perched premiere bicycle singing world
	Mistral: exiting scientist her‡ speaking a‡	**Mistral**: lady else beauty her‡ hands	**Mistral**: makeup accessories getting her shopping
women	**SNLI**: ladies† woman‡ performing a‡ group	**SNLI**: woman‡ party a‡ group tall	**SNLI**: lunch men† they a‡ play
	GPT-4: ladies‡ females‡ lady‡ conversation walking	**GPT-4**: ladies‡ fruits vegetables female‡ restaurant	**GPT-4**: suits‡ ladies† men‡ meeting business
	Llama: costumes gathering sporting dancing socializing	**Llama**: ladies shopping choreographed store local	**Llama**: men‡ football celebrities‡ during competing
	Mistral: athletes people‡ clothing street outdoors	**Mistral**: females‡ female‡ woman‡ singing show	**Mistral**: people‡ clothing being any performers

Table 6: Top-five words in hypotheses of a particular label by PMI with gender-related query words in the premise, filtered to co-occurrences of at least three. (Hypothesis words that also appear in the premise are not included.) Significance of a likelihood ratio test for independence denoted by † ($\alpha = 0.01$) and ‡ ($\alpha = 0.001$).

premise words are often associated with *suit, man, men, football, meeting, competing, business*, which might demonstrate a gender bias.

Race & age biases Unlike gender-related query terms, race and age-related query terms (e.g. african, asian, elderly, old) yield unclear stereotypical associations. For most race or ethnicity premise words, the words with the highest PMI were uninformative, e.g. *is, the,* and *a*. For age-related queries, the most associated words in entailed hypotheses were synonyms (*senior, older*), and in contradictions were antonyms (*young, children*.)

Gender-related stereotypical associations seem stronger than racial and ethnic biases in LLM-generated datasets. One possible explanation is that LLM-generated hypotheses typically mention racial and ethnicity-related words much less often than in SNLI's hypotheses, as shown in Figure 5.[8]

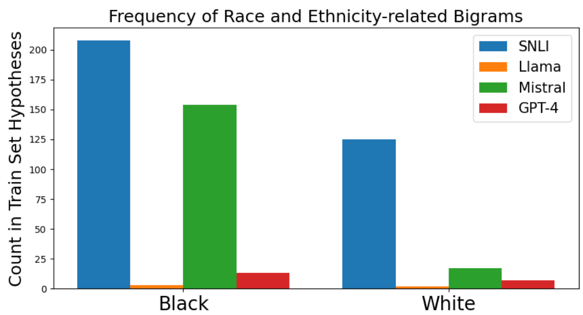


Figure 5: Number of hypotheses in each train set that contain race-related words followed by one of the people-related words from Rudinger et al. (2017).

6 Conclusion

We studied whether Natural Language Inference datasets created by eliciting hypotheses from LLMs contain biases. We used 3 LLMs to recreate a por-

[8]In the figure, "black" refers to the words *black* and *african*, "white" refers to the words *white* and *european*. The people-related words are the person-related query words from Rudinger et al. (2017): *woman, man, women, men, girl, boy, girls, boys, female, male, mother, father, sister, brother, daughter, son, person* and *people*.

tion of SNLI and applied standard techniques to determine that like SNLI, LLM-elicited datasets contain annotation artifacts and stereotypical biases. On our LLM-generated NLI datasets, fine-tuned BERT hypothesis-only classifiers achieve between 86-96% accuracy. Our analyses indicated that LLMs rely on similar strategies and heuristics as crowd-source workers when creating entailed, neutral, and contradicted hypotheses in response to a premise. Our results provide further empirical evidence that well-attested biases in human-elicited text persist in LLM-generated text. Our findings provide a cautionary tale for relying on unfiltered, out-of-the-box LLM-generated textual data for NLP datasets.

7 Limitations

Srikanth and Rudinger (2022) showed that while NLI models *can* gain high performance while ignoring the premise, in practice models still condition on the premise context when making predictions. While our work demonstrated that LLM-elicited datasets can contain biases, it is unclear to what extent these biases harm NLI model robustness.

While we aimed to mirror the process used to generate SNLI, our approach is not perfectly comparable. First, SNLI was created by a large pool of crowd-source workers while we focus on just 3 LLMs. Secondly, crowd-source workers could ask clarifying questions, but LLMs could not. Thirdly, the one-shot nature of our prompting prevented LLMs from incorporating instructions across premises, such as the FAQ suggestion to not "[reuse] the same sentence."

Another limitation of our work is that we relied on a single prompt to elicit hypotheses from LLMs. Recent work has demonstrated that seemingly insignificant changes to prompts can result in widely varying responses (Mizrahi et al., 2024). We leave a multi-prompt analysis for future work.

Acknowledgments

We thank anonymous reviewers from current and past versions of the article for their insightful comments and suggestions. This research benefited from support by an internal grant awarded by the Bryn Mawr College Faculty Awards and Grants Committee.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Hritik Bansal, Karthik Gopalakrishnan, Saket Dingliwal, Sravan Bodapati, Katrin Kirchhoff, and Dan Roth. 2023. Rethinking the role of scale for in-context learning: An interpretability-based case study at 66 billion scale. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11833–11856, Toronto, Canada. Association for Computational Linguistics.

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *arXiv preprint arXiv:2306.15766*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Zheng Cai, Lifu Tu, and Kevin Gimpel. 2017. Pay attention to the ending:strong neural baselines for the ROC story cloze task. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 616–622, Vancouver, Canada. Association for Computational Linguistics.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023. DISCO: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada. Association for Computational Linguistics.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, page 177–190, Berlin, Heidelberg. Springer-Verlag.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Shi Feng, Eric Wallace, and Jordan Boyd-Graber. 2019. Misleading failures of partial-input baselines. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5533–5538, Florence, Italy. Association for Computational Linguistics.

Johann Frei and Frank Kramer. 2023. Annotated dataset creation through large language models for non-english medical nlp. *Journal of Biomedical Informatics*, page 104478.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

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.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 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)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Himanshu Gupta, Kevin Scaria, Ujjwala Anantheswaran, Shreyas Verma, Mihir Parmar, Saurabh Arjun Sawant, Chitta Baral, and Swaroop Mishra. 2024. TarGEN: Targeted data generation with large language models. In *First Conference on Language Modeling*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2024. Generative language models exhibit social identity biases. *Nature Computational Science*, pages 1–11.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *Preprint*, arXiv:2307.16883.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Skylar Kolisko and Carolyn Jane Anderson. 2023. Exploring social biases of large language models in a college artificial intelligence course. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15825–15833.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Katelyn Mei, Sonia Fereidooni, and Aylin Caliskan. 2023. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1699–1710, New York, NY, USA. Association for Computing Machinery.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. volume 12, pages 933–949, Cambridge, MA. MIT Press.

Anders Giovanni Møller, Arianna Pera, Jacob Dalsgaard, and Luca Aiello. 2024. The parrot dilemma: Human-labeled vs. LLM-augmented data in classification tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 179–192, St. Julian's, Malta. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation. *arXiv preprint arXiv:2306.00176*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *Preprint*, arXiv:2304.03277.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. 2024. Breaking bias, building bridges: Evaluation and mitigation of social biases in llms via contact hypothesis. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1180–1189.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 15–25, Vancouver, Canada. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.

Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong Park. 2024. Ask LLMs directly, "what shapes your bias?": Measuring social bias in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16122–16143, Bangkok, Thailand. Association for Computational Linguistics.

Neha Srikanth and Rachel Rudinger. 2022. Partial-input baselines show that NLI models can ignore context, but they don't. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4753–4763, Seattle, United States. Association for Computational Linguistics.

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. *Preprint*, arXiv:2307.09288.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*.

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

Ben Wang and Aran Komatsuzaki. 2021. Gpt-j-6b: A 6 billion parameter autoregressive language model.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

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

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

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Canwen Xu, Corby Rosset, Ethan Chau, Luciano Corro, Shweti Mahajan, Julian McAuley, Jennifer Neville, Ahmed Awadallah, and Nikhil Rao. 2024a. Automatic pair construction for contrastive post-training. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 149–162, Mexico City, Mexico. Association for Computational Linguistics.

Ran Xu, Hejie Cui, Yue Yu, Xuan Kan, Wenqi Shi, Yuchen Zhuang, May Dongmei Wang, Wei Jin, Joyce Ho, and Carl Yang. 2024b. Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15496–15523, Bangkok, Thailand. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Eyal Shnarch, and Leshem Choshen. 2024. Achieving human parity in content-grounded datasets generation. In *The Twelfth International Conference on Learning Representations*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Mengjie Zhao, Fei Mi, Yasheng Wang, Minglei Li, Xin Jiang, Qun Liu, and Hinrich Schuetze. 2022. LMTurk: Few-shot learners as crowdsourcing workers in a language-model-as-a-service framework. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 675–692, Seattle, United States. Association for Computational Linguistics.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, pages 1–55.

A Appendix

Query	ENTAIL	NEUTRAL	CONTRA
african	**SNLI**: are is the	**SNLI**: a to are the is	**SNLI**: a are is the
	GPT-4: kids a are is person	**GPT-4**: group of performing a are	**GPT-4**: a are playing man swimming
	Llama: a are is people person	**Llama**: his at are a to	**Llama**: man group a playing are
	Mistral: an people a in are	**Mistral**: an or be may are	**Mistral**: an not and are is
asian	**SNLI**: an† for with near the	**SNLI**: chinese work up an waiting	**SNLI**: american‡ white black taking from
	GPT-4: city cooking having food woman	**GPT-4**: sushi lunch tourists busy exploring	**GPT-4**: party a‡ men dancing child
	Llama: students shopping city a‡ food	**Llama**: cultural individual class restaurant heading	**Llama**: models† they astronauts preparing shoot
	Mistral: individual‡ women outdoor an‡ area	**Mistral**: exploring city tourists‡ an‡ collecting	**Mistral**: an‡ green outside their cars
asians	**SNLI**: are the	**SNLI**: are the	**SNLI**: are the
	GPT-4: people are	**GPT-4**: of	**GPT-4**: park are
	Llama: dining‡ people are	**Llama**: of	**Llama**: the are
	Mistral: asian‡ are	**Mistral**: asian‡ are	**Mistral**: asian‡ are
caucasian	**SNLI**: white‡ is	**SNLI**: is	**SNLI**: the
	GPT-4: is	**GPT-4**: is	**GPT-4**: man is swimming
	Llama: is	**Llama**: is	**Llama**: is
	Mistral: is	**Mistral**: the is	**Mistral**: not is the
chinese	**SNLI**: is	**SNLI**: is the	**SNLI**: a
	GPT-4: are is	**GPT-4**: a in is	**GPT-4**: a is in
	Llama: a is	**Llama**: someone are is	**Llama**: a is
	Mistral: is there	**Mistral**: a be the	**Mistral**: not is the
indian	**SNLI**: the is	**SNLI**: a to the is	**SNLI**: on is the
	GPT-4: people is are	**GPT-4**: is	**GPT-4**: a is pool swimming
	Llama: a people is are person	**Llama**: is	**Llama**: group a in is
	Mistral: an people is are there	**Mistral**: a the is	**Mistral**: the on are is

Table 7: Race, Ethnicity, and Nationality-Related Queries

Query	ENTAIL	NEUTRAL	CONTRA
elderly	**SNLI**: old‡ an‡ wearing a are	**SNLI**: old he a‡ an is	**SNLI**: old a man at is
	GPT-4: old‡ senior‡ citizen lady instrument	**GPT-4**: senior‡ old‡ jazz festival musician	**GPT-4**: young‡ children a playing man
	Llama: an‡ instrument musical for a	**Llama**: seniors‡ older‡ citizen‡ senior‡ an	**Llama**: young‡ child concert woman fashion
	Mistral: seniors‡ older† an‡ for the	**Mistral**: older music an‡ a of	**Mistral**: an‡ a playing is on
old	**SNLI**: elderly‡ not a‡ an person	**SNLI**: hair just home out an	**SNLI**: young‡ has two a people
	GPT-4: elderly‡ gentleman‡ citizen‡ senior‡ something	**GPT-4**: citizens† grandson‡ citizen‡ elderly‡ grandmother‡	**GPT-4**: young‡ sandbox her a‡ girl
	Llama: produce† elderly‡ woman an† resting	**Llama**: elderly‡ citizen‡ senior‡ grandfather an‡	**Llama**: young‡ children child‡ her toy
	Mistral: elderly‡ an‡ woman walking a	**Mistral**: older‡ elderly grandmother† grandson grandfather	**Mistral**: elderly‡ young‡ woman an a‡
teenagers	**SNLI**: are the	**SNLI**: are	**SNLI**: are the
	GPT-4: young‡ outside people are	**GPT-4**: high students school game group†	**GPT-4**: children‡ library playing are pool
	Llama: activity engaging people in are	**Llama**: group of friends are a	**Llama**: are the
	Mistral: children young people are there	**Mistral**: could it be are	**Mistral**: are not the
young	**SNLI**: off building jumps a‡ he	**SNLI**: alone funny high brothers beach	**SNLI**: kite books birds practicing swims
	GPT-4: children‡ activities physical child‡ a‡	**GPT-4**: teenagers test cap giant teenager‡	**GPT-4**: snowman adults‡ teenagers old rocking
	Llama: feature kids‡ sunny observing creative	**Llama**: teenagers‡ mom skatepark weekend games	**Llama**: nursing† seniors‡ citizens elderly senior
	Mistral: shore studying acrobatics children‡ sandy	**Mistral**: females learning skills siblings school	**Mistral**: pants kids‡ they toys a‡

Table 8: Age-Related Queries

Query	ENTAIL	NEUTRAL	CONTRA
boy	**SNLI**: boys† child a‡ his is	**SNLI**: boys a‡ down trying his	**SNLI**: girl‡ up asleep a‡ nobody
	GPT-4: active his playground male trick	**GPT-4**: hide seek kid‡ teenager swimming	**GPT-4**: kid his‡ girl‡ classroom quietly
	Llama: child‡ a‡ urban enjoying playing	**Llama**: young‡ summer person a‡ kid	**Llama**: surfing teenager suit tie working
	Mistral: a‡ young‡ group child† standing	**Mistral**: child‡ how young‡ practicing swimming	**Mistral**: a‡ man subject‡ reading is
boys	**SNLI**: playing are the	**SNLI**: their and of are a	**SNLI**: girls‡ playing are† the
	GPT-4: children‡ sport event activity participating	**GPT-4**: kids‡ game their playing group	**GPT-4**: are‡ beach a swimming the
	Llama: children‡ physical activity† engaging† outdoors	**Llama**: sport kids‡ team participating game	**Llama**: players competing team game astronauts
	Mistral: children‡ event sport outdoors playing	**Mistral**: children‡ sport running kids‡ fun	**Mistral**: kids‡ photo inside individuals in
girl	**SNLI**: girls her a‡ child wearing	**SNLI**: girls she a‡ plays her†	**SNLI**: she guy boy a‡ wearing
	GPT-4: female‡ a‡ riding musical instrument	**GPT-4**: woman‡ young‡ teenager lady‡ child	**GPT-4**: boy‡ video a‡ his climbing
	Llama: a‡ wearing place public the	**Llama**: instrument expressing woman‡ young‡ favorite	**Llama**: ice child‡ professional mountain toy
	Mistral: wearing a‡ young physical activity	**Mistral**: woman‡ young‡ subject little her	**Mistral**: a‡ any wearing subject book
girls	**SNLI**: girl some‡ their wearing are	**SNLI**: girl some they at are	**SNLI**: boys‡ their two playing a
	GPT-4: females‡ children‡ game sport participating	**GPT-4**: match group team a‡ practicing	**GPT-4**: boys‡ field studying football soccer
	Llama: students athletes indoors activity physical	**Llama**: teenagers† teammates women† friendly sisters	**Llama**: celebrities‡ premiere cats movie show
	Mistral: sports celebrating people‡ are‡ there†	**Mistral**: females‡ female children† athletes could†	**Mistral**: children individuals‡ a park are
female	**SNLI**: woman‡ a is the	**SNLI**: woman‡ wearing a in is	**SNLI**: male‡ woman playing a is
	GPT-4: woman‡ athlete the playing performing	**GPT-4**: woman† practicing lady her a	**GPT-4**: skiing basketball mountain man a
	Llama: a playing person is	**Llama**: woman‡ a of is	**Llama**: fashion man playing a is
	Mistral: woman‡ performing an playing is	**Mistral**: exercise woman a be for	**Mistral**: a subject playing person is
he	**SNLI**: man a†	**SNLI**: a	**SNLI**: man a
	GPT-4: man‡ wearing a† in person	**GPT-4**: in his a	**GPT-4**: a pool in swimming
	Llama: wearing a in person	**Llama**: in for a the	**Llama**: cooking pool swimming at a
	Mistral: a in person	**Mistral**: someone a for be	**Mistral**: a person not
male	**SNLI**: man a people outside is	**SNLI**: practicing man‡ from his a	**SNLI**: waiting an man his sitting
	GPT-4: man‡ at performing a two	**GPT-4**: man† a† park at on	**GPT-4**: skiing a mountain woman cooking
	Llama: their a outdoors on is	**Llama**: man‡ practicing break couple on	**Llama**: sunny preparing man an park
	Mistral: man performing space riding a	**Mistral**: man‡ a† performing his couple	**Mistral**: subject a wearing bench sitting

Table 9: Additional Gender-Related Queries

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo. *Example: For the caption “Two dogs are running through a field.” you could write “There are animals outdoors."*
- Write one alternate caption that **might be** a **true** description of the photo. *Example: For the caption “Two dogs are running through a field." you could write “Some puppies are running to catch a stick."*
- Write one alternate caption that is **definitely** a **false** description of the photo. *Example: For the caption “Two dogs are running through a field." you could write “The pets are sitting on a couch." This is different from the* maybe correct *category because it’s impossible for the dogs to be both running and sitting.*

In response to the original caption, please return the 3 alternate captions in a JSON readable format and include no other commentary.

Here is an example of the correct format of response to the prompt:
Original caption: "Two dogs are running through a field"
Three JSON-parseable alternate captions, with "definitely true", "might be true", and "definitely false" descriptions of the photo:
{"true": "There are animals outdoors.",
"maybe": "Some puppies are running to catch a stick.",
"false": "The pets are sitting on a couch." }

Now, please generate the 3 alternate captions following the JSON-parseable format described earlier:
Original Caption: **[INSERT SNLI PREMISE]**
Three JSON-parseable alternate captions, with "definitely true", "might be true", and "definitely false" descriptions of the photo:

Figure 6: The prompt provided to all LLMs. The first four paragraphs are identical to those provided to MTurk workers for the SNLI dataset.