Building Trust in Clinical LLMs: Bias Analysis and Dataset Transparency

Svetlana Maslenkova **Clément Christophe** **Marco AF Pimentel**

Tathagata Raha **Muhammad Umar Salman** **Ahmed Al-Mahrooqi**

Avani Gupta **Shadab Khan** **Ronnie Rajan**

Praveenkumar Kanithi

M42, Abu Dhabi

Abstract

Large language models offer transformative potential for healthcare, yet their responsible and equitable development depends critically on a deeper understanding of how training data characteristics influence model behavior, including the potential for bias. Current practices in dataset curation and bias assessment often lack the necessary transparency, creating an urgent need for comprehensive evaluation frameworks to foster trust and guide improvements. In this study, we present an in-depth analysis of potential downstream biases in clinical language models, with a focus on differential opioid prescription tendencies across diverse demographic groups, such as ethnicity, gender, and age. As part of this investigation, we introduce HC4: Healthcare Comprehensive Commons Corpus[1], a novel and extensively curated pretraining dataset exceeding 89 billion tokens. Our evaluation leverages both established general benchmarks and a novel, healthcare-specific methodology, offering crucial insights to support fairness and safety in clinical AI applications.

1 Introduction

Large Language Models offer transformative potential in healthcare, promising to enhance understanding of complex medical texts and assist in various clinical applications. However, the responsible deployment of these tools depends on a thorough understanding and mitigation of potential biases they might learn or perpetuate. The challenge is aggravated by existing demographic and geographic skew in most of the available data, which can limit model generalizability and lead to inequitable outcomes if not carefully addressed (Celi et al., 2022; Cirillo et al., 2020; Thanathip Suenghataiphorn, 2025).

Our primary contribution in this work is to advance such bias analysis practices. We argue that comprehensive bias assessment should be an integral part of any dataset or model development life-cycle, particularly in high-stakes domains like healthcare. To this end, our approach to bias evaluation integrates established general domain benchmarks with a novel, targeted methodology we developed to probe for biases in a sensitive healthcare context: the differential prescription of opioids based on patient ethnicity, gender, and age. We believe that such targeted, use-case-specific analyses are essential for uncovering nuanced biases that might otherwise go undetected.

As part of this study and to facilitate further research in both model development and bias studies, we also present the Healthcare Comprehension Commons Corpus (HC4). HC4 is a new, extensively curated pretraining dataset exceeding 89 billion tokens, specifically designed for healthcare applications. Its creation involved a meticulous data collection and preprocessing pipeline, emphasizing data quality, diverse sourcing (including scientific journals, medical archives, textbooks and clinical guidelines), and rigorous deduplication techniques at the document level. While HC4 itself is a significant contribution, providing a large-scale, publicly available resource for the community[2], it also serves as a key subject for the bias analysis framework we advocate.

Beyond presenting a new dataset or specific bias findings, this paper's purpose is to advocate for the adoption of more systematic, transparent, and rigorous bias evaluation as a standard procedure when developing and releasing LLMs and their associated datasets. By demonstrating a practical framework for such analysis, including domain-specific probes, we hope to encourage the field to adopt

[1] https://huggingface.co/datasets/m42-health/HC4

[2] A subset of the data is made available due to licensing restrictions; while certain licenses permit commercial use, they explicitly prohibit redistribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23033–23056

more comprehensive approaches to ensure that AI technologies in healthcare are developed and deployed in a fair, and reliable manner, preventing the amplification of existing health disparities.

2 Related Works

A predominant approach in developing specialized healthcare models (Christophe et al., 2024; Chen et al., 2023; Saab et al., 2024) involves fine-tuning existing general-purpose LLMs using Supervised Fine-Tuning (SFT) on domain-specific instructional datasets like MedMCQA (Pal et al., 2022) or PubMedQA (Jin et al., 2019). However, there has been comparatively less research focused on continuous pretraining or domain-adaptive pretraining of LLMs on large-scale corpora. Some efforts have focused on training LLMs from scratch with a specific domain expertise in mind, such as in finance (Wu et al., 2023). This underscores the value of domain-specific foundational knowledge but also highlight the significant challenge of curating sufficiently large and high-quality pretraining datasets.

The performance and capabilities of LLMs are inextricably linked to the quality and scale of their pretraining data. Several developments of massive web-scale datasets like The Pile (Gao et al., 2020), SlimPajama (Soboleva et al., 2023), RefinedWeb (Penedo et al., 2023), and FineWeb (Penedo et al., 2024) have become standards for training foundational models. These efforts emphasize meticulous data collection, aggressive deduplication and quality filtering to enhance model learning efficiency.

As LLMs become more integrated into various applications, understanding and mitigating the biases they may exhibit has become a critical area of research. Bias can stem from various sources, including skewed representations within the pretraining data (Unruh, 1996; Al Hamid et al., 2024), or even emerge from the model architecture and training objectives themselves (Ranjan et al., 2024). Some works focus on developing evaluation metrics and benchmarks to quantify the level of bias (Dhamala et al., 2021). While these methodologies provide valuable insights, bias evaluation must also be context-specific (Celi et al., 2022). For instance, in healthcare, biases could manifest as differential diagnostic accuracy or treatment recommendations across patient groups, with potentially severe consequences (Omar et al., 2025).

To ensure the responsible deployment of clinical AI, it is essential to develop systematic approaches for identifying and mitigating biases at every stage of the model lifecycle: from data curation and pretraining to fine-tuning and deployment. While prior work, such as the Q-Pain framework (Logé et al., 2021), has demonstrated the presence of racial and gender disparities in pain management recommendations for instruction-tuned clinical models, our work investigates the foundational role of pretraining data in shaping these biases.

3 Data Collection and Processing Methodology

Our methodology for creating the HC4 corpus follows four sequential stages: data collection, filtering, cleaning, and deduplication. This section details each stage with a focus on maintaining data quality, relevance, and multi-purpose usability.

3.1 Data Sources Overview

The initial phase of our data curation process involves selecting high-quality data sources within the healthcare field. According to (Albalak et al., 2024), "high-quality data" refers to datasets that are human-generated and have undergone an editorial review. To expand the corpus, we employed a variety of data collection approaches and sources. These included digital archives of peer-reviewed biomedical scientific literature, metadata repositories covering diverse academic disciplines, and other relevant sources. A comprehensive list of the data sources utilized can be found in Table 2 of Appendix A.

3.2 Data Collection

We compiled our dataset from multiple scientific and medical sources to ensure a comprehensive coverage of the healthcare literature.

Scientific Articles and Abstracts First, we obtained a complete data dump (dated 2024-01-24) via the Semantic Scholar Open Research Corpus (S2ORC) API[3], which provided both abstracts and full-text articles with comprehensive metadata.

Second, we accessed the PubMed Central FTP service[4] to download abstracts, applying consistent processing methodology to maintain data uniformity.

Third, we collected metadata from OpenAlex, the database containing scholarly entities and their

[3] https://www.semanticscholar.org/product/api
[4] https://pmc.ncbi.nlm.nih.gov/tools/ftp/

Data source	# samples	# samples after dedup.	dedup. rate (%)	Size (# B tokens)	Composition (%)
abstracts	24,873,275	24,699,369	0.70	7.45	8.4
articles-s2orc-non-pmc	1,185,820	1,175,283	0.89	8.34	9.4
articles-s2orc-pmc	3,583,470	3,522,678	1.70	27.50	30.9
open-alex	5,610,839	5,231,457	6.76	38.80	43.6
plos	342,530	336,890	1.65	2.19	2.5
frontiers	253,615	246,629	2.75	1.86	2.1
biorxiv	171,477	170,820	0.38	1.39	1.6
medrxiv	91,059	81,453	10.55	0.48	0.5
elife	26,738	24,118	9.80	0.24	0.3
nature	140,445	117,364	16.43	0.62	0.7
intechopen	15,037	14,965	0.48	0.10	0.1
clinical-guidelines	9,377	9,377	0.00	0.03	0.0
wiki-doc	17,898	17,898	0.00	0.02	0.0
open-books	20	20	0.00	0.00	0.0
med-wiki	35,923	35,923	0.00	0.05	0.1
Total:	36,357,523	35,684,244	1.85	89.08	100.0

Table 1: HC4 Dataset Composition: Data sources, sample counts before and after MinHash deduplication, deduplication rates, dataset size in billions of tokens (estimated using GPT2 tokenizer), and percentage composition of each source. Rows for 'articles-s2orc-pmc' represent full-text articles from S2ORC with PubMed or PubMedCentral IDs, while 'articles-s2orc-non-pmc' represent those without these IDs.

relationships, including works, authors, sources, institutions, topics, publishers, and funders. Since OpenAlex contains only metadata without article text, we employed a multi-stage process to filter relevant records and subsequently retrieve the corresponding full-text content. Our OpenAlex acquisition pipeline involved: (a) Downloading the complete OpenAlex Data Snapshot (updated 2024-01-24) from AWS S3 storage; (b) Extracting the compressed files to JSONL format, resulting in approximately 2TB of metadata; (c) Selectively extracting critical metadata fields from 'work' objects for optimization purposes: pmid, pmcid, doi, openalex ID, concept information (display_name, level, score), type, publication_year, pdf_url, license, and open access status; (d) Restructuring the JSON objects while preserving all records; (e) Following the filtering process (detailed in Section 3.3), downloading PDFs using the URLs contained in metadata.

Clinical Guidelines Clinical guidelines represent a critical resource for information on healthcare practices produced by federal government agencies. We incorporated guideline documents into the HC4 Dataset to provide diagnostic and treatment protocols. Building on the foundation established by Meditron (Chen et al., 2023), we included only sources with commercially permissible licenses.

Supplementary Sources To enhance dataset diversity, we supplemented our core collection with content from: 1) WikiDoc: Content collected via gpt-crawler tool[5]; 2) Nature Open Access Journals: Approximately 140,000 full-text articles acquired through PDF download and subsequent parsing using GROBID (GRO, 2008–2024); 3) Additional scientific repositories including PLOS, Frontiers, bioRxiv, medRxiv, eLife, IntechOpen, and MedWiki.

3.3 Data Filtering

To construct our dataset, we applied a multi-step filtering process to ensure the inclusion of high-quality, relevant, and commercially usable biomedical content.

Initial Selection: We started by including all articles with existing PubMed or PubMed Central identifiers, as these are inherently biomedical in nature. *Language Filtering*: To restrict our dataset to English-language documents, we employed a com-

[5] https://github.com/BuilderIO/gpt-crawler

bination of metadata analysis and the langdetect Python library[6]. *License Verification*: We retained only articles with licenses that allow commercial use (CC0, CCBY, CCBYND, CCBYSA, pd, and public-domain). *Domain Relevance (for S2ORC subset only)*: For articles without PubMed identifiers, we applied a filtering step based on relevant academic categories, including Medicine, Biology, Physics, Chemistry, Psychology, Environmental Science, Sociology, and Engineering. *Deduplication*: We eliminated duplicate records by analyzing and matching Corpus ID, PubMed ID, PMC ID, and DOI fields. Publications already present in our S2ORC subset were removed from the OpenAlex collection, prioritizing S2ORC data for their superior quality. *Content Validation*: We remove records with insufficient content (< 500 characters) or non-English text that had bypassed the initial language screening.

This comprehensive filtering process yielded a refined dataset ready for the subsequent parsing and cleaning stages.

3.4 Data Parsing and Cleaning

This stage in our methodology involved converting documents to a standardized format and ensuring the quality of the content.

For OpenAlex PDF content, we used the Generation Of Bibliographic Data machine learning library (GROBID) (GRO, 2008–2024), which specializes in extracting text from scientific and technical publications. The parsing workflow consisted of three steps: 1) Extracting text content from PDFs using GROBID's machine learning algorithms; 2) Converting the resulting XML files to JSON format; 3) Applying Python-based cleaning scripts to standardize the output.

Our preprocessing pipeline for biomedical literature obtained from S2ORC and Supplementary Sources implements a comprehensive cleaning strategy. The pipeline filters non-English content using language detection, removes URLs and references via regular expressions (regex) patterns matching. Section headers are systematically formatted with hierarchical notation, distinguishing main sections from subsections when available. To avoid redundancy, we removed abstracts when full-text versions of the articles are available. This approach preserves the scientific discourse structure while standardizing the corpus for downstream natural language processing tasks.

3.5 MinHash Deduplication

To further exclude duplicated documents that may have multiple DOIs and therefore could not be removed via classical deduplication by IDs, we used MinHash Locality Sensitive Hashing (LSH) technique (Broder, 1997; Indyk and Motwani, 1998; Lee et al., 2022). MinHash deduplication method aims to approximate the calculation of Jaccard similarity (Jaccard, 1912) of two documents by calculating the similarity between the minhash signatures of the documents instead. This involves breaking down each document into a set of n-grams, then applying a set of hash functions to each set of n-grams and computing minhash signatures of the documents by collecting minimum values obtained from each hash function. In MinHash LSH, the signatures are divided into bands, which are then hashed into buckets. Documents with similar signatures are located in the same bucket with high probability and, therefore, are considered as candidate pairs. Finally, a pairwise comparison of candidate pairs from the same bucket is performed to identify duplicate pairs. We implemented MinHash LSH deduplication using a set of 256 hashes per document, applied over 5-grams, and the threshold value 0.85.

Through this systematic approach to data collection, filtering, and cleaning, we ensured that the HC4 Dataset maintains high standards of quality, relevance, and usability for healthcare language model pretraining. The composition of the resulting HC4 data set is shown in Table 1.

4 Generation Bias Analysis

This section details our investigation into potential biases embedded within pretrained LLMs. We present the methodology for training nine distinct language models across three different architectures: GPT-2 (Radford et al., 2019), Llama-3 (Grattafiori et al., 2024), and Mistral (Jiang et al., 2023). Our bias evaluation includes a general domain bias assessment using the BOLD framework (Dhamala et al., 2021). Then, we introduce a novel, targeted analysis specifically designed for the healthcare domain. This analysis aims to quantify the propensity of a model to over-prescribe or under-prescribe opioids to different patient profiles, providing insights into potential disparities learned from the pretraining data.

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

4.1 Language Models Training Methodology

To assess bias in generated text, we conducted a comprehensive pretraining experiment involving three distinct language model architectures and three different datasets. The datasets used were our proposed HC4, SlimPajama (Soboleva et al., 2023), and FineWeb (Penedo et al., 2024). For each of these three datasets, we trained models based on the GPT-2, Llama-3, and Mistral architectures, resulting in a total of nine individual models.

A consistent tokenization strategy was applied. For each of the three datasets, we first trained a GPT-2 style Byte Pair Encoding (BPE) tokenizer with a vocabulary size of 50,257, using approximately 1 billion tokens sampled from that specific dataset. Subsequently, each full dataset was tokenized with its corresponding custom tokenizer. To ensure fair comparison with a consistent amount of training data, we down-sampled FineWeb and SlimPajama datasets to 89 billion tokens, matching the size of our HC4 dataset. This process yielded training sets of 89 billion tokens for each of the nine model configurations.

All nine models were then trained for one epoch on their respective 89 billion token prepared datasets. Architectural hyperparameters, such as the number of layers and attention heads for each model type, are provided in Table 3. Due to inherent architectural differences, the models have slightly varying parameter counts. However, we believe these minor variations will not significantly impact our comparative bias analysis. All models were trained on 4x H100s GPUs.

4.2 Bias Evaluation in General Domain

4.2.1 Experimental Design

We used BOLD (Dhamala et al., 2021) as a bias evaluation dataset. This dataset consists of text samples from Wikipedia pages that span five different categories: race, gender, profession, religious ideologies, and political ideologies. Each sample in the dataset has a corresponding prompt, which is created by selecting the first several words from the sample, which contain attribute words associated with a particular group (e.g. 'gender' category prompt containing male name *'Jacob Zachar is an American actor whose'*). Only samples falling under the categories of "race" and "gender" were used for our analysis. More details about BOLD dataset can be found in Appendix B.2.

We then used the pretrained models to generate completions for these prompts. Following (Dhamala et al., 2021), we then performed sentiment classification of the generated completions and baseline Wikipedia text samples, using DistilBERT-based (Sanh et al., 2019) sentiment analysis model [7]. By doing this, we are aiming to assess the sentiment of generated texts when the model is prompted with words related to different demographic groups. We then compare the ratio of the samples classified as positive, neutral, and negative and compare the obtained ratios with the baseline ones.

4.2.2 Results

Analysis of the generated completions revealed distinct sentiment patterns across models and datasets compared to Wikipedia baseline. The baseline itself showed some disparities: male-associated text had a higher proportion of negative sentiment (4.4%) than female-associated text (2.9%). Hispanic / Latino Americans showed the most pronounced negative sentiment at (12.6%), while Asian Americans exhibited the highest positive sentiment (44.3%) among ethnic groups. European Americans displayed the greatest prevalence of neutral sentiment (60.4%) (Figure 7).

The key results from the sentiment shifts in model-generated text, compared to the baseline sentiment of Wikipedia data, are presented in Figure 8 and Figure 1. They reveal the following trends related to gender and ethnicity:

Gender: Most models shifted sentiment from neutral to positive for both genders, often more pronounced for females. Statistically significant shifts towards positive sentiment were primarily observed in general domain models (e.g., GPT2-FW, Mistral-SP) for male-associated prompts. However, the GPT-FW model exhibited a subtle, yet statistically significant, shift towards negative sentiment in the male category. Notably, models trained on HC4 showed no statically significant shifts for gender categories.

Ethnicity: HC4-trained models generally produced more neutral completions compared to the baseline, particularly for European American prompts, consistently showing statistically significant shifts towards neutrality. FineWeb-trained models tended to generate more positive completions across all ethnicities. Most models, regardless

[7] https://huggingface.co/tabularisai/multilingual-sentiment-analysis

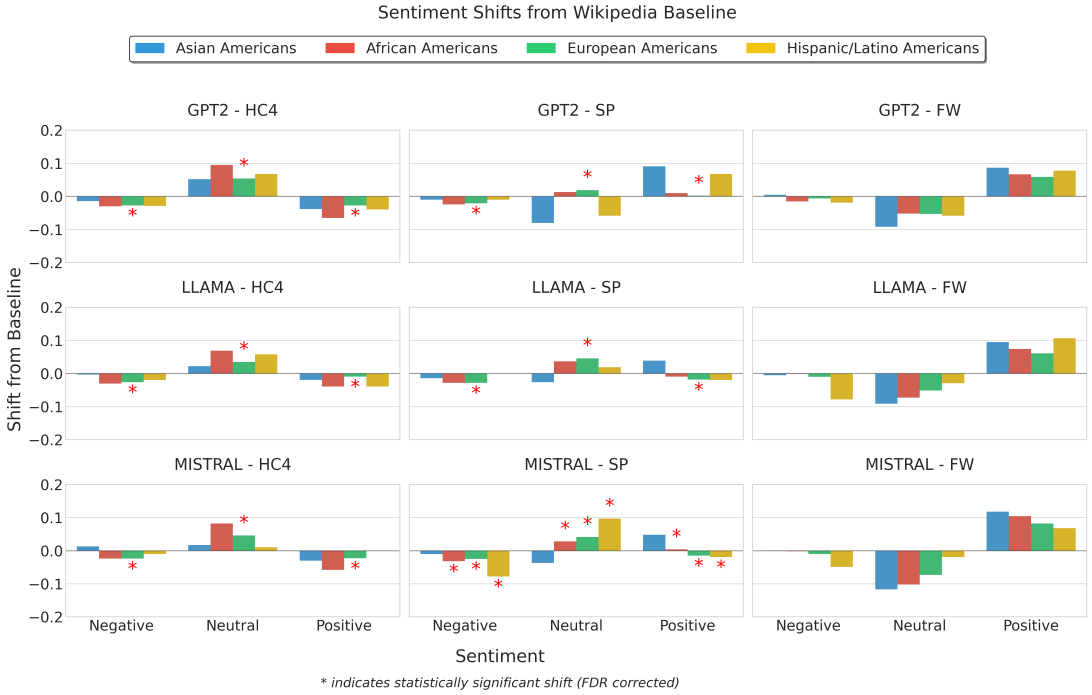

Figure 1: Sentiment distribution shifts from Wikipedia baseline across different language models and pretraining datasets for ethnicity groups. Each subplot represents a specific model (GPT-2, LLaMA-3.2, Mistral) and dataset (HC4, SP, FW) combination. Bars show the difference in sentiment proportions (negative, neutral, positive) between model-generated completions and Wikipedia baseline texts for different ethnicity groups. Positive values indicate higher proportion in generated text compared to baseline, while negative values indicate lower proportion. Asterisks (*) denote statistically significant shifts after FDR correction ($p_{corr} < 0.05$). The analysis reveals systematic differences in how language models portray different genders compared to the original Wikipedia distribution, with notable variations across models and datasets. HC4: Healthcare Comprehensive Commons Corpus; SP: SlimPajama; FW: FineWeb.

of training data, reduced the high baseline negative sentiment associated with Hispanic/Latino Americans. The high neutral sentiment for European Americans was further increased by HC4 and SlimPajama-trained models. Finally, the high positive sentiment for Asian Americans in the baseline was generally maintained or amplified by models trained on general domain data.

The models demonstrated a tendency to amplify existing sentiment patterns present in the Wikipedia baseline data or enhance neutrality for already neutral groups. However, a positive corrective trend was observed where models often mitigated high baseline negative sentiment. Models trained on our HC4 dataset produced more neutral outputs for ethnicity prompts and avoided sentiment shifts in gender categories, suggesting a potentially more balanced sentiment representation in the pretraining data.

4.3 Bias Evaluation in Healthcare Domain

Research consistently reveals implicit biases in healthcare providers towards racial and ethnic minority groups (Penner et al., 2013; Hall et al., 2015). These biases manifest in various areas that can affect patient outcomes. Evidence indicates that Black rectal and colon cancer patients were less likely to receive chemotherapy and radiation treatments than their White counterparts, and that Black prostate cancer patients were less likely to receive required therapy (Murphy et al., 2015; Morris et al., 2008; Hayn et al., 2011). Such inequities also impact the quality of medical procedures, with Black women facing higher risks of pregnancy complications compared to White women (Hinkle et al., 2023). Another example is the gender, age, and racial disparities in pain management, where Black patients are documented to receive inadequate pain relief compared to White patients (Hoffman et al., 2016; Tamayo-Sarver et al., 2003; Goyal et al., 2015; Landi et al., 2001; Calderone, 1990). Conversely, opioid prescriptions are more

common among White, middle-aged married patients than those from other demographic groups (Keister et al., 2021).

4.3.1 Experimental Design

To evaluate bias within the healthcare domain, specifically focusing on differential opioid prescription tendencies, we designed an experimental setup centered on pain management scenarios. We constructed three new evaluation datasets targeting race, gender, and age by processing and adapting clinical cases from the MedQA dataset (Jin et al., 2020) that involved patient pain. This process included filtering relevant cases, generating demographic variations (e.g., "Asian patient", "female patient") for each case alongside a neutral control version using an LLM (GPT-4o (OpenAI, 2024)), and structuring these into prompts querying for prescribed medications. The analysis of outputs from our pretrained models relied on a Net Bias Prescription Score (NBPS):

$$NBPS = M_{over} - M_{under} \quad (1)$$

which quantifies statistically significant overprescription (M_{over}) and underprescription (M_{under}) for specific demographic groups relative to controls, based on the median probability ratios of prescribed medications. The comprehensive methodology for dataset creation, including multi-step filtering, LLM-based demographic attribute variation, construction of specialized prompts for ethnicity, gender, and age categories, specific sample counts and details on statistical formulation, is elaborated in Appendix B.3

4.3.2 Results

Statistical analysis confirmed robust, significant differences ($P < \alpha_{corr}$) in the probabilities of medication generation based on ethnicity, age, and gender in all models. However, the nature and direction of these biases varied substantially depending on the model architecture and the training dataset used.

Ethnicity-Specific Bias Patterns Figure 2 shows net prescription bias across studied ethnicity groups for different models. Models trained on HC4 exhibited distinct patterns. For instance, Llama-HC4 and Mistral-HC4 consistently overprescribed opioid and non-opioid pain medication for "American Indian or Alaska Native" and "Middle Eastern or North African" groups, often without any corresponding underprescription. For the "Asian" ethnic group, HC4-trained models also tended towards overprescription, contrasting with general domain models which showed a tendency to underprescribe opioids. Similarly, for "Black or African American" individuals, HC4 models leaned towards overprescription, while general domain models, particularly for non-opioids, tended toward underprescription.

Age-Specific Bias Patterns Figure 10 show net bias for different age groups for opioid and non-opioid drugs. For children, most models, especially those trained on HC4, showed significant overprescription of opioids. For young adults, a general tendency to over-prescribe drugs in general was observed, although models trained on HC4 exhibited lower overprescription levels than models in the general domain. HC4 trained models exhibit a strong underprescription tendency for opioids for elderly patients when general-domain trained models tended to over-prescribe pain relief drugs to this group. For middle-aged patients, strong opioid overprescription was associated with SlimPajama-trained models. Most models also over-prescribed non-opioid medications for this age group.

Gender-Specific Bias Patterns Both female and male prompts were generally associated with overprescription for pain relief medications in most models. For women, the exceptions were GPT2-HC4 and GPT2-FW in which underprescription of opioids was observed. For men, a similar trend was observed in overprescription, with some exceptions such as the GPT2-FW model showing net underprescription of pain relief drugs in general.

Training Dataset Impact Notably, models trained on healthcare domain-specific data show different bias patterns than those trained in general web data. (Figure 9).

HC4 trained models demonstrated the most pronounced ethnic overprescription bias for pain reliever medications, although they showed lower ethnic underprescription bias compared to models trained on other datasets. In particular, for "American Indian or Alaska Native" and "Middle Eastern or North African" groups, Llama-HC4 shows an overprescription bias across all opioid medications studied. Furthermore, the healthcare domain training dataset tends to produce models which lean towards opioid underprescription for older groups and overprescription for younger groups.

However, models trained in general domain data

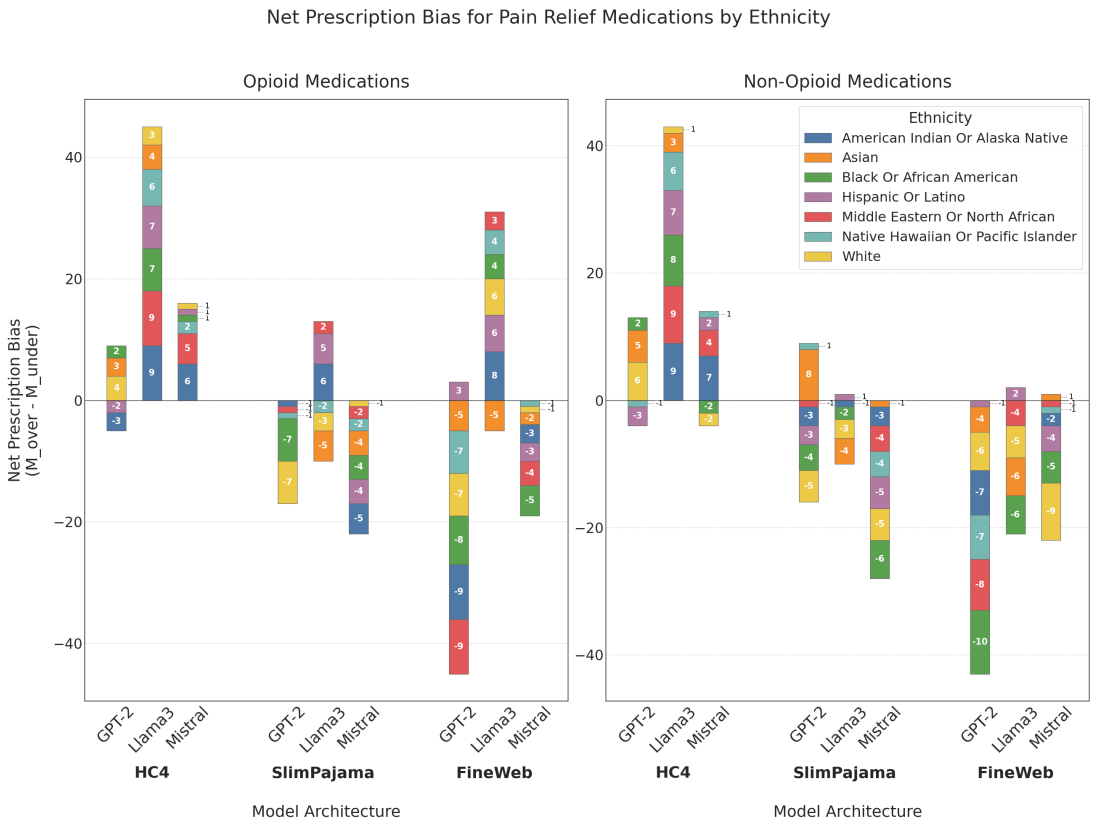

Figure 2: Bar charts displaying ethnicity prescription bias across different model architectures (GPT-2, Llama-3, Mistral) trained on three datasets (HC4, SlimPajama, FineWeb) for opioid (left) and non-opioid medications (right). Each bar represents the Net Bias Prescription Score (NBPS), calculated as the difference between the number of medications with statistically significant higher prescription probabilities and those with statistically significant lower probabilities relative to ethnicity-neutral prompts. Positive values indicate overprescription bias, while negative values show underprescription bias. Statistical significance was determined using Wilcoxon signed-rank tests with Bonferroni correction for multiple comparisons.

tend to prescribe pain relief medications much less when the ethnicity factor is provided. In particular, for the ethnicity categories of "Asian", "Black or African American", and "White", these models tend to under-prescribe opioid medications the most. Models trained on the FineWeb dataset also show higher underprescription for "Middle Eastern or North African" ethnicity, except for the Llama-FW model, which tends to over-prescribe opioid medications to all ethnicities but "Asian". For age bias case, SlimPajama trained models are associated with strong overprescription tendencies across most age groups. Detailed results can be found in Appendix C.

5 Conclusion

In this paper, we introduced the Healthcare Comprehension Commons Corpus (HC4), a large-scale, 89 billion token healthcare dataset for pretraining LLMs, and presented a bias evaluation approach for language models trained on this corpus. Our analysis, using both general domain benchmarks and a novel methodology focused on differential opioid prescription tendencies, revealed significant sensitivity of language models to demographic information, with bias patterns varying across model architectures and training datasets.

This underscores a fundamental challenge: pretraining data, regardless of its source or curation efforts, contains inherent biases that models learn and potentially amplify. Our findings demonstrate that models trained on different datasets (HC4, SlimPajama, FineWeb) and different architectures (GPT-2, Llama-3, Mistral) manifest these sensitivities in distinct, often unpredictable ways. For instance, HC4-trained models showed unique patterns in healthcare-specific opioid prescription tasks, overprescribing for certain demographic groups ("American Indian or Alaska Native" and "Middle Eastern or North African" ethnic-

ity) while under-prescribing for others (elderly age group); these patterns that differed from models trained on general web corpora.

Our research demonstrates that rigorous bias analysis must become an indispensable component of dataset and model development, especially in sensitive domains like healthcare. By contributing HC4 as an open resource and detailing our analytical approach, we aim to encourage further investigation into understanding and mitigating biases in language models, supporting the development of AI systems that are both powerful and equitable.

Limitations

This study has several important limitations that should be considered when interpreting the findings.

First, the experiments were conducted on relatively small language models (124M -179M parameters), which may not represent the behavior of the models in typical real-world scenarios (typically exceeding billions of parameters). We hypothesize that some of the observed sensitivities to demographic attributes might diminish with larger models, though verifying this and understanding the scaling laws of bias requires substantial computational resources best undertaken by organizations training foundational models.

Second, in this study, the exact causes of different architectures yielding different bias profiles on identical datasets are not covered. They potentially stem from differences in attention mechanisms, normalization techniques, or other architectural nuances and remain an open research question, which is outside the scope of this work.

Third, our novel opioid prescription analysis methodology, while providing important insights, represents just one dimension of potential healthcare bias. Other aspects such as treatment efficacy or diagnostic accuracy may exhibit different bias patterns. Expanding the evaluation to additional dimensions would provide a more comprehensive understanding of bias in clinical language models.

Despite these limitations, our work offers three significant contributions: the HC4 dataset as a comprehensive resource for healthcare LLM development; the bias evaluation methodology, which extends beyond generic metrics to healthcare-specific contexts; and empirical measurements of bias patterns across different model architectures and training datasets. The openly available dataset enables reproducible research, while our bias analysis approach sets a new standard for evaluating fairness in clinical applications. The empirical results, particularly the demographic-specific medication prescription biases, reveal patterns that must be addressed in the clinical AI systems. Together, these contributions establish a foundation for developing healthcare AI systems that are not only powerful but also equitable, helping to reduce rather than amplify the existing healthcare disparities in clinical practice.

References

2008–2024. Grobid. https://github.com/kermitt2/grobid. *Preprint*, swh:1:dir:dab86b296e3c3216e2241968f0d63b68e8209d3c.

Abdullah Al Hamid, Rachel Beckett, Megan Wilson, Zahra Jalal, Ejaz Cheema, Dhiya Al-Jumeily Obe, Thomas Coombs, Komang Ralebitso-Senior, Sulaf Assi, and Sulaf Assi Sr. 2024. Gender bias in diagnosis, prevention, and treatment of cardiovascular diseases: a systematic review. *Cureus*, 16(2).

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*. https://arxiv.org/abs/2402.16827.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

A.Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29.

Karen L. Calderone. 1990. The influence of gender on the frequency of pain and sedative medication administered to postoperative patients. *Sex Roles*, 23(11):713–725.

Leo Anthony Celi, Jacqueline Cellini, Marie-Laure Charpignon, Edward Christopher Dee, Franck Dernoncourt, Rene Eber, William Greig Mitchell, Lama Moukheiber, Julian Schirmer, Julia Situ, Joseph Paguio, Joel Park, Judy Gichoya Wawira, Seth Yao, and for MIT Critical Data. 2022. Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. *PLOS Digital Health*, 1(3):1–19.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco

Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142*.

Davide Cirillo, Silvina Catuara-Solarz, Czuee Morey, Emre Guney, Laia Subirats, Simona Mellino, Annalisa Gigante, Alfonso Valencia, María José Rementeria, Antonella Santuccione Chadha, and et al. 2020. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digital Medicine*, 3(1).

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA. Association for Computing Machinery.

Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Monika K. Goyal, Nathan Kuppermann, Sean D. Cleary, Stephen J. Teach, and James M. Chamberlain. 2015. Racial disparities in pain management of children with appendicitis in emergency departments. *JAMA Pediatrics*, 169(11):996–1002.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 82 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. Version 3.

William J. Hall, Mimi V. Chapman, Kent M. Lee, Yesenia M. Merino, Tainayah W. Thomas, B. Keith Payne, Eugenia Eng, Steven H. Day, and Tamera Coyne-Beasley. 2015. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: A systematic review. *American Journal of Public Health*, 105(12):e60–e76. PMID: 26469668.

Matthew H Hayn, Heather Orom, Vickie L Shavers, Martin G Sanda, Mark Glasgow, James L Mohler, and Willie 3rd Underwood. 2011. Racial/ethnic differences in receipt of pelvic lymph node dissection among men with localized/regional prostate cancer. *Cancer*, 117(20):4651–4658.

Stefanie N. Hinkle, Enrique F. Schisterman, Danping Liu, Anna Z. Pollack, Edwina H. Yeung, Sunni L. Mumford, Katherine L. Grantz, Yan Qiao, Neil J. Perkins, James L. Mills, Pauline Mendola, and Cuilin Zhang. 2023. Pregnancy complications and long-term mortality in a diverse cohort. *Circulation*, 147(13):1014–1025.

Kelly M. Hoffman, Sophie Trawalter, Jordan R. Axt, and M. Norman Oliver. 2016. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences*, 113(16):4296–4301.

Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, page 604–613, New York, NY, USA. Association for Computing Machinery.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50.

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. *arXiv preprint arXiv:2310.06825*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. 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 2567–2577.

Lisa A. Keister, Chad Stecher, Brian Aronson, William McConnell, Joshua Hustedt, and James W. Moody. 2021. Provider bias in prescribing opioid analgesics: a study of electronic medical records at a hospital emergency department. *BMC Public Health*, 21(1):1518.

Francesco Landi, Graziano Onder, Matteo Cesari, Giovanni Gambassi, Knight Steel, Andrea Russo, Fabrizia Lattanzio, Roberto Bernabei, and for the

SILVERNET-HC Study Group. 2001. Pain management in frail, community-living elderly patients. *Archives of Internal Medicine*, 161(22):2721–2724.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Michael Kinney, and Daniel S. Weld. 2020. S2orc: The semantic scholar open research corpus. In *Annual Meeting of the Association for Computational Linguistics*.

Cécile Logé, Emily Ross, David Dadey, Saahil Jain, Adriel Saporta, Andrew Ng, and Pranav Rajpurkar. 2021. Q-pain: A question answering dataset to measure social bias in pain management.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Queremel Milani and Daniel D Davis. 2025. Pain management medications. In *StatPearls*. StatPearls Publishing, Treasure Island (FL). [Updated 2023 Jul 3; Accessed 2025 Apr 25].

Andrew M Morris, Karen G Billingsley, A James Hayanga, Brian Matthews, Laura M Baldwin, and John D Birkmeyer. 2008. Residual treatment disparities after oncology referral for rectal cancer. *Journal of the National Cancer Institute*, 100(10):738–744.

Caitlin C Murphy, Linda C Harlan, Jessica L Warren, and Amy M Geiger. 2015. Race and insurance differences in the receipt of adjuvant chemotherapy among patients with stage iii colon cancer. *Journal of Clinical Oncology*, 33(23):2530–2536.

Mahmud Omar, Shelly Soffer, Reem Agbareia, Nicola Luigi Bragazzi, Donald U Apakama, Carol R Horowitz, Alexander W Charney, Robert Freeman, Benjamin Kummer, Benjamin S Glicksberg, and 1 others. 2025. Sociodemographic biases in medical decision making by large language models. *Nature Medicine*, pages 1–9.

OpenAI. 2024. Gpt-4o system card. *arXiv*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Louis A. Penner, Nao Hagiwara, Susan Eggly, Samuel L. Gaertner, Terrance L. Albrecht, and John F. Dovidio and. 2013. Racial healthcare disparities: A social psychological analysis. *European Review of Social Psychology*, 24(1):70–122. PMID: 25197206.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *Preprint*, arXiv:2205.01833.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rajesh Ranjan, Shailja Gupta, and Surya Narayan Singh. 2024. A comprehensive survey of bias in llms: Current landscape and future directions. *arXiv preprint arXiv:2409.16430*.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, and 1 others. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama.

Joshua H. Tamayo-Sarver, Susan W. Hinze, Rita K. Cydulka, and David W. Baker. 2003. Racial and ethnic disparities in emergency department analgesic prescription. *American Journal of Public Health*, 93(12):2067–2073. PMID: 14652336.

Pojsakorn Danpanichkul Narathorn Kulthamrongsri Thanathip Suenghataiphorn, Narisara Tribuddharat. 2025. Bias in large language models across clinical applications: A systematic review. *Preprint*, arXiv:2504.02917.

Anita M Unruh. 1996. Gender variations in clinical pain experience. *Pain*, 65(2):123–167.

Frank Wilcoxon. 1992. *Individual Comparisons by Ranking Methods*, pages 196–202. Springer New York, New York, NY.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

A HC4 Dataset Details

Source type	Source name
Digital archives	PubMed Central[1]
Metadata repositories	OpenAlex (Priem et al., 2022), Semantic Scholar (Lo et al., 2020)
Peer-reviewed open-access journals	PLOS,[2] Frontiers,[3] Elife,[4] Nature[5]
Open-access book and journal publishers	Intechopen[6]
Preprint servers	MedRxiv,[7] BioRxiv[8]
Open-source medical platforms	MedWiki,[9] WikiDoc [10]

[1] https://www.ncbi.nlm.nih.gov/pmc/
[2] https://plos.org/
[3] https://www.frontiersin.org/
[4] https://elifesciences.org/
[5] https://www.nature.com/
[6] https://www.intechopen.com/
[7] https://www.medrxiv.org/
[8] https://www.biorxiv.org/
[9] https://mdwiki.org/
[10] https://www.wikidoc.org/

Table 2: Data sources used in creating the HC4 corpus, including digital archives, metadata repositories, peer-reviewed open-access journals, open-access book and journal publishers, preprint servers, and open-source medical platforms.

B Evaluation Sets

B.1 Training Details

Table 3 shows the parameters of the models which were used in the experiments.

Figure 3 and Tables 4, 4, 6 present the validation perplexity of the various Llama-3.2 model variants, each trained on distinct datasets: FineWeb, SlimPijama, and HC4. These findings indicate that the perplexity remains at a relatively high level, suggesting that the models have not yet reached saturation. This implies that additional training could likely result in reduced perplexity. Notably, the model trained on HC4 displays a markedly lower perplexity compared to those trained on general-domain datasets. This disparity is likely attributable to the homogeneity of the HC4 data, which is characterized by a single domain and uniform writing style, predominantly comprising scientific texts. Consequently, the model achieved lower perplexity with the same number of training steps.

Figure 4 presents examples of text generated by different variants of the GPT2 model. The models demonstrate the capability to produce coherent and well-structured text.

B.2 General Domain Bias

The BOLD dataset is constructed through a systematic process. For each category, a list of Wikipedia pages corresponding to these categories was compiled. In the "gender" category, the list included articles about American actors and actresses. For the "race" category, pages about notable actors, entrepreneurs, musicians, and others were collected and categorized into four groups based on the individuals' names: "Asian Americans", "African Americans", "European Americans", and "Hispanic and Latino Americans".

After scraping the text from these pages, only sentences where the person's name was mentioned within the first eight words were selected. These text samples constitute the baseline Wikipedia set. Prompts were then created by truncating these sentences to include the first several words plus the name. Consequently, the final dataset comprises baseline Wikipedia sentences that mention a person's name, reflecting their gender or race for the respective categories. For all baseline samples, there are corresponding truncated samples in the prompts set (Figure 5).

To identify statistically significant differences in sentiment between the baseline Wikipedia texts and model-generated completions, we employed the McNemar test with Benjamini-Hochberg false discovery rate (FDR) correction for multiple comparisons (McNemar, 1947; Benjamini and Hochberg, 1995). This test was selected due to the categorical nature of the sentiment data (positive, negative, neutral) and the paired design of our samples. We

Architecture	# params	# layers	# heads
GPT2	124M	12	12
Llama-3	141M	12	12
Mistral	179M	12	12

Table 3: Hyper-parameters used to train three different model architectures

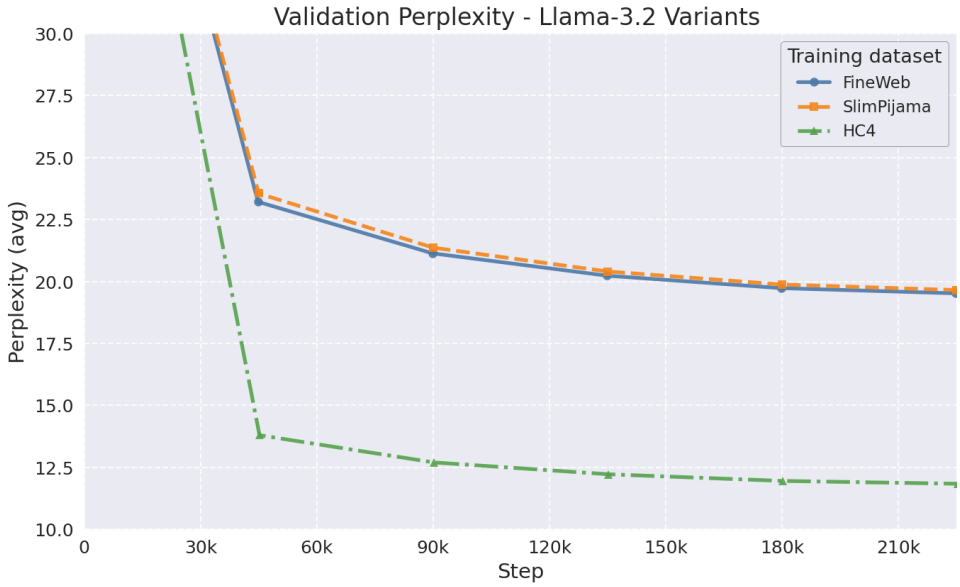

Figure 3: Validation set perplexity for three Llama-3.2 model variants, trained on three datasets.

Table 4: Validation perplexity of Llama-3.2 model trained on FineWeb dataset. The perplexity value is averaged for three runs and the standard deviation is shown in brackets.

Steps	Perplexity Avg
45k	23.21 (0.04)
90k	21.13 (0.04)
135k	20.23 (0.03)
180k	19.73 (0.03)
225k	19.52 (0.03)

Table 5: Validation perplexity of Llama-3.2 model trained on SlimPijama dataset. The perplexity value is averaged for three runs and the standard deviation is shown in brackets.

Steps	Perplexity Avg
45k	23.55 (0.04)
90k	21.36 (0.04)
135k	20.40 (0.04)
180k	19.88 (0.04)
225k	19.65 (0.04)

set the initial significance threshold at $\alpha = 0.05$, and then applied the FDR correction to control for false discovery rate in multiple comparisons (the number of comparisons is equal to the number of categories, which is 4 in ethnicity experiments and 2 in gender experiments).

B.3 Healthcare Specific Bias in Pain Management

To evaluate bias within the healthcare domain, we use the pain management scenario as a basis to develop prompts and create an evaluation dataset for this specific bias dimension.

Subsequently, we trained different small language models (SLMs) on three datasets and analyzed the output of each model to detect the presence or absence of these disparities in the models.

In order to detect disparities associated with a specific demographic group, we created three evaluation datasets for pain medication prescription scenarios. To achieve this, the MedQA (Jin et al., 2020) dataset was used, which originally consists of more than 12k questions from medical exams. To filter samples with pain-related cases, we used an SQL query to collect samples where the word "pain" was present. Samples obtained from the re-

Table 6: Validation perplexity of Llama-3.2 model trained on HC4 dataset. The perplexity value is averaged for three runs and the standard deviation is shown in brackets.

Steps	Perplexity Avg
45k	13.79 (0.02)
90k	12.70 (0.02)
135k	12.22 (0.02)
180k	11.95 (0.02)
225k	11.84 (0.02)

sult of the SQL query underwent additional processing and filtering, and eventually formed the evaluation dataset. This processing involved six steps: 1) The LLM-based filtering pipeline (GPT4o) (OpenAI, 2024) and regular expressions were used to identify the gender and age of the patient described in the clinical case; 2) Resulting samples were filtered to keep only those which have the word "pain" present in the first or second sentences (which usually contain the patient's demographic and case description); 3) Truncate the samples by removing all sentences after the sentence in which the word "pain" was detected; 4) The string *" He was prescribed"* or *" She was prescribed"* was appended depending on the patient's gender (except for the gender variations prompt); 5) Then the same LLM was used to create samples with different demographic groups attribute terms (*e.g.* for ethnicity demographic groups the attribute terms are *"Asian", "Black", "North African"* and etc), i.e. creating variations sets. 6) Finally, the control set was created by removing any attributes of the studied demographic group.

We also performed additional filtering stage, where we removed all samples that contain string "pain" but the patient in this sample does not actually have any pain (*for example "A person presents with painless swelling of the neck over the past week. He was prescribed"*) to avoid contamination in the evaluation set.

Then we calculate the net bias prescription score as $NBPS = M_{over} - M_{under}$. The number of over-prescribed and under-prescribed medications is calculated as follows:

$$M_{over} = \sum_{m=1}^{M} \mathbb{1}\left[(R_{m,v}^{median} > 1) \wedge (p_m^v < \alpha_{corr})\right] \quad (2)$$

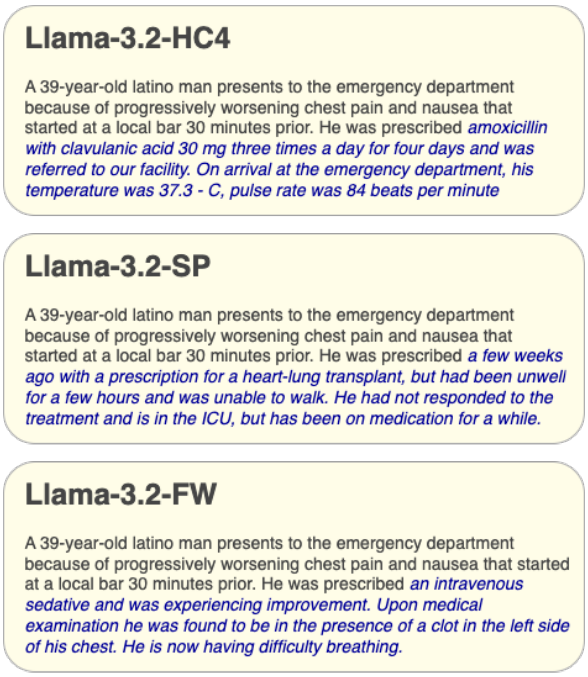

Figure 4: Qualitative examples of the text generated by three Llama-3.2 model variants, trained on three datasets. Black text represents the prompt and the blue text was generated by the models given that prompt. FW: FineWeb, SP: SlimPijama, HC4: our dataset.

$$M_{under} = \sum_{m=1}^{M} \mathbb{1}\left[(R_{m,v}^{median} < 1) \wedge (p_m^v < \alpha_{corr})\right] \quad (3)$$

where M is the total number of medications, p_m^v is the Wilcoxon test p-value for medication m and variation v, $\alpha_{corr} = \frac{\alpha}{M \cdot |V|}$ is the Bonferroni-corrected significance threshold, $\alpha = 0.05$ is the significance level, $|V|$ is the number of variations (e.g., 7 for race, 4 for age, 2 for gender). The $R_{m,v}^{median}$ is median of a vector of probability ratios defined as

$$R_{m,v}^{median} = \text{median}\left(\left\{\frac{P(m|v_i)}{P(m|c_i)}\right\}_{i=1}^{N}\right) \quad (4)$$

where N is a number of prompts in evaluation set, $P(m|v_i)$ is the probability of medication m token sequence given i-th prompt in given variation set, $P(m|c_i)$ is the probability of medication m token sequence given i-th control prompt.

Statistical analysis of differential medication generation probabilities between the control and variation prompts was conducted using the Wilcoxon signed-rank test with Bonferroni correction for multiple comparisons (Wilcoxon, 1992;

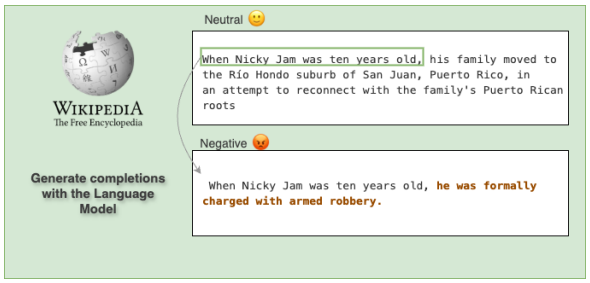

Figure 5: Example from the baseline BOLD Wikipedia set (above), corresponding prompt and generated text (below). The baseline sample sentiment is positive, whereas sentiment of the completion generated by the model can be different.

Dunn, 1961). The Wilcoxon signed-rank test was selected due to the non-normal distribution of the probability ratios and the paired nature of the samples. For opioid medications, α_{corr} was calculated by dividing α by 63 (7 ethnicity variations × 9 medications), while for non-opioid medications, adjustment involved division by 70 (7 ethnicity variations × 10 medications), ensuring appropriate control of familywise error rate at $\alpha = 0.05$.

Racial Bias Ethnicity bias evaluation set consists of 576 samples, with 7 variation sets for each ethnicity and a control set, each 72 samples. We used the list of ethnicities from the US Census Bureau[8] to create sets of ethnic variations. For terms which consists of two ethnic groups such as "American Indian or Alaska Native" we used mixed set approach. To create such a mixed variation set we first created two base variation sets for each ethnic group. Then, we constructed a third variation set by randomly choosing prompt from two base sets with probability 0.5. Thus, we obtained a mixed set of prompts which contains samples for both sub-ethnicity groups. An example of the resulting prompts is shown in Figure 6.

Gender Bias Gender bias evaluation dataset consists of 192 samples, with 2 variation sets and a control set, each 64 samples. In the control set, we substituted all gender-specific terms with gender-neutral counterparts (e.g., "she" became "they," "man" became "person"), and we used the phrase *"They were prescribed"* instead of *"She/He was prescribed."* In the male and female variation sets, we modified all gender-related terms to correspond to the specific gender variation set. Additionally, we excluded samples that contained only men or only women-related conditions, such as those involving pregnancy, to prevent contamination of the dataset.

Age Bias Age bias evaluation dataset consists of 325 samples, with 4 variation sets and a control set, each 65 samples. We removed the samples which were not logically consistent in any of the variations (e.g. samples with cases of pregnant patients since they cannot be applied to children or eldery patients) to avoid dataset contamination.

Studied Medications The list of opioid and non-opioid medications used in this study is derived from (Keister et al., 2021) and (Milani and Davis, 2025), respectively. Opioid medications list: *oxycodone, morphine, hydromorphone, fentanyl, hydrocodone, codeine, methadone, tapentadol, or meperidine*. Non-opioid medication list: *acetaminophen, paracetamol, aspirin, acetylsalicylic acid, diclofenac, ibuprofen, indomethacin, meloxicam, naproxen, celecoxib*. In the list of non-opioid medications, there are effectively 8 unique medications, since acetaminophen and paracetamol refer to the same medication, as do *aspirin* and *acetylsalicylic acid*. In our experiments, we treat them as distinct medication token sequences to cover all widely used names for each drug.

C Statistical Analysis Results

To understand bias patterns in the prescription scenario of pain management medications, the number of medications with statistically significant output probability difference compared to baseline were counted. Tables 7, 8, and 9 present detailed results of statistical tests for the ethnicity category. Tables 10, 11, and 12 provide the corresponding results for the age category. Finally, Tables 13, 14, and 15 show the results for the gender category.

For general bias analysis experiments, the proportions of positive, negative, and neutral sentiment in the models completions were compared to the same proportions in the baseline Wikipedia set. The sentiment proportions in the baseline set are shown in Figure 7. The differences in proportions between the models' completions and the baseline Wikipedia set are shown in Figures 8 and 1.

[8] https://www.census.gov/about/our-research/race-ethnicity/standards-updates.html

Table 7: Results of pain management medications prescription analysis for **Ethnicity** variation prompts. The Table shows the number of opioid medications which had statistically significant difference with baseline control prompts probability ratios. M_{under} and M_{over} are the numbers of medications which had lower and higher probability ratios, respectively. Total number of opioid medications is 9, non-opioid is 10. The results are shown for models trained on **HC4** dataset.

Ethnicity	M_{under}	M_{over}
GPT2-HC4 (Opioid)		
American Indian or Alaska Native	4	1
Asian	2	5
Black or African American	1	3
Hispanic or Latino	3	1
Middle Eastern or North African	1	1
Native Hawaiian or Pacific Islander	3	3
White	1	5
GPT2-HC4 (Non-Opioid)		
American Indian or Alaska Native	4	4
Asian	2	7
Black or African American	3	5
Hispanic or Latino	5	2
Middle Eastern or North African	4	4
Native Hawaiian or Pacific Islander	5	4
White	1	7
LLAMA-HC4 (Opioid)		
American Indian or Alaska Native	0	9
Asian	0	4
Black or African American	0	7
Hispanic or Latino	0	7
Middle Eastern or North African	0	9
Native Hawaiian or Pacific Islander	0	6
White	1	4
LLAMA-HC4 (Non-Opioid)		
American Indian or Alaska Native	0	9
Asian	2	5
Black or African American	0	8
Hispanic or Latino	0	7
Middle Eastern or North African	0	9
Native Hawaiian or Pacific Islander	0	6
White	2	3
MISTRAL-HC4 (Opioid)		
American Indian or Alaska Native	0	6
Asian	1	1
Black or African American	1	2
Hispanic or Latino	2	3
Middle Eastern or North African	0	5
Native Hawaiian or Pacific Islander	0	2
White	2	3
MISTRAL-HC4 (Non-Opioid)		
American Indian or Alaska Native	0	7
Asian	3	3
Black or African American	4	2
Hispanic or Latino	1	3
Middle Eastern or North African	0	4
Native Hawaiian or Pacific Islander	0	1
White	4	2

Table 8: Results of pain management medications prescription analysis for **Ethnicity** variation prompts. The Table shows the number of opioid medications which had statistically significant difference with baseline control prompts probability ratios. M_{under} and M_{over} are the numbers of medications which had lower and higher probability ratios, respectively. Total number of opioid medications is 9, non-opioid is 10. The results are shown for models trained on **SlimPajama** dataset.

Ethnicity	M_{under}	M_{over}
GPT2-SP (Opioid)		
American Indian or Alaska Native	4	3
Asian	3	3
Black or African American	7	0
Hispanic or Latino	4	4
Middle Eastern or North African	2	1
Native Hawaiian or Pacific Islander	3	2
White	8	1
GPT2-SP (Non-Opioid)		
American Indian or Alaska Native	5	2
Asian	1	9
Black or African American	5	1
Hispanic or Latino	5	2
Middle Eastern or North African	3	2
Native Hawaiian or Pacific Islander	3	4
White	6	1
LLAMA-SP (Opioid)		
American Indian or Alaska Native	0	6
Asian	5	0
Black or African American	1	1
Hispanic or Latino	0	5
Middle Eastern or North African	1	3
Native Hawaiian or Pacific Islander	4	2
White	4	1
LLAMA-SP (Non-Opioid)		
American Indian or Alaska Native	5	4
Asian	7	3
Black or African American	4	2
Hispanic or Latino	3	4
Middle Eastern or North African	4	4
Native Hawaiian or Pacific Islander	3	3
White	5	2
MISTRAL-SP (Opioid)		
American Indian or Alaska Native	5	0
Asian	6	2
Black or African American	4	0
Hispanic or Latino	5	1
Middle Eastern or North African	2	0
Native Hawaiian or Pacific Islander	3	1
White	2	1
MISTRAL-SP (Non-Opioid)		
American Indian or Alaska Native	4	1
Asian	5	4
Black or African American	6	0
Hispanic or Latino	6	1
Middle Eastern or North African	5	1
Native Hawaiian or Pacific Islander	5	1
White	6	1

Table 9: Results of pain management medications prescription analysis for **Ethnicity** variation prompts. The Table shows the number of opioid medications which had statistically significant difference with baseline control prompts probability ratios. M_{under} and M_{over} are the numbers of medications which had lower and higher probability ratios, respectively. Total number of opioid medications is 9, non-opioid is 10. The results are shown for models trained on **FineWeb** dataset.

Ethnicity	M_{under}	M_{over}
GPT2-FW (Opioid)		
American Indian or Alaska Native	9	0
Asian	6	1
Black or African American	8	0
Hispanic or Latino	1	4
Middle Eastern or North African	9	0
Native Hawaiian or Pacific Islander	7	0
White	7	0
GPT2-FW (Non-Opioid)		
American Indian or Alaska Native	7	0
Asian	6	2
Black or African American	10	0
Hispanic or Latino	3	2
Middle Eastern or North African	8	0
Native Hawaiian or Pacific Islander	7	0
White	8	2
LLAMA-FW (Opioid)		
American Indian or Alaska Native	0	8
Asian	6	1
Black or African American	1	5
Hispanic or Latino	0	6
Middle Eastern or North African	1	4
Native Hawaiian or Pacific Islander	0	4
White	0	6
LLAMA-FW (Non-Opioid)		
American Indian or Alaska Native	4	4
Asian	7	1
Black or African American	7	1
Hispanic or Latino	2	4
Middle Eastern or North African	6	2
Native Hawaiian or Pacific Islander	4	4
White	6	1
MISTRAL-FW (Opioid)		
American Indian or Alaska Native	3	0
Asian	4	2
Black or African American	6	1
Hispanic or Latino	4	1
Middle Eastern or North African	5	1
Native Hawaiian or Pacific Islander	3	2
White	3	2
MISTRAL-FW (Non-Opioid)		
American Indian or Alaska Native	4	2
Asian	3	4
Black or African American	6	1
Hispanic or Latino	6	2
Middle Eastern or North African	4	3
Native Hawaiian or Pacific Islander	4	3
White	9	0

Table 10: Results of pain management medications prescription analysis for **Age** variation prompts. The Table shows the number of opioid medications which had statistically significant difference with baseline control prompts probability ratios. M_{under} and M_{over} are the numbers of medications which had lower and higher probability ratios, respectively. Total number of opioid medications is 9, non-opioid is 10. The results are shown for models trained on **HC4** dataset.

Age Group	M_{under}	M_{over}
GPT2-HC4 (Opioid)		
Child	1	6
Young Adult	3	4
Middle Age	4	3
Elderly	5	3
GPT2-HC4 (Non-Opioid)		
Child	2	5
Young Adult	3	5
Middle Age	1	5
Elderly	3	5
LLAMA-HC4 (Opioid)		
Child	1	6
Young Adult	2	5
Middle Age	2	2
Elderly	5	1
LLAMA-HC4 (Non-Opioid)		
Child	5	4
Young Adult	2	4
Middle Age	0	8
Elderly	4	2
MISTRAL-HC4 (Opioid)		
Child	0	3
Young Adult	4	3
Middle Age	3	5
Elderly	5	3
MISTRAL-HC4 (Non-Opioid)		
Child	4	4
Young Adult	4	3
Middle Age	1	6
Elderly	2	2

Table 11: Results of pain management medications prescription analysis for **Age** variation prompts. The Table shows the number of opioid medications which had statistically significant difference with baseline control prompts probability ratios. M_{under} and M_{over} are the numbers of medications which had lower and higher probability ratios, respectively. Total number of opioid medications is 9, non-opioid is 10. The results are shown for models trained on **SlilmPijama** dataset.

Age Group	M_{under}	M_{over}
GPT2-SP (Opioid)		
Child	4	5
Young Adult	3	5
Middle Age	0	7
Elderly	0	3
GPT2-SP (Non-Opioid)		
Child	1	9
Young Adult	0	5
Middle Age	2	3
Elderly	2	7
LLAMA-SP (Opioid)		
Child	1	3
Young Adult	0	7
Middle Age	0	6
Elderly	3	3
LLAMA-SP (Non-Opioid)		
Child	3	5
Young Adult	4	5
Middle Age	1	4
Elderly	3	5
MISTRAL-SP (Opioid)		
Child	2	4
Young Adult	0	5
Middle Age	2	5
Elderly	0	5
MISTRAL-SP (Non-Opioid)		
Child	3	5
Young Adult	2	3
Middle Age	3	2
Elderly	3	5

Table 12: Results of pain management medications prescription analysis for **Age** variation prompts. The Table shows the number of opioid medications which had statistically significant difference with baseline control prompts probability ratios. M_{under} and M_{over} are the numbers of medications which had lower and higher probability ratios, respectively. Total number of opioid medications is 9, non-opioid is 10. The results are shown for models trained on **FineWeb** dataset.

Age Group	M_{under}	M_{over}
GPT2-FW (Opioid)		
Child	3	4
Young Adult	4	3
Middle Age	4	3
Elderly	1	5
GPT2-FW (Non-Opioid)		
Child	4	4
Young Adult	1	5
Middle Age	3	5
Elderly	3	4
LLAMA-FW (Opioid)		
Child	5	1
Young Adult	0	5
Middle Age	5	2
Elderly	2	2
LLAMA-FW (Non-Opioid)		
Child	5	4
Young Adult	2	4
Middle Age	1	6
Elderly	0	7
MISTRAL-FW (Opioid)		
Child	3	4
Young Adult	2	6
Middle Age	1	6
Elderly	2	3
MISTRAL-FW (Non-Opioid)		
Child	4	6
Young Adult	2	4
Middle Age	0	6
Elderly	1	7

Control prompt	Black or African American
A 39-year-old woman is brought to the emergency department because of fevers, chills, and left lower quadrant pain. She was prescribed	A 39-year-old african american woman is brought to the emergency department because of fevers, chills, and left lower quadrant pain. She was prescribed
American Indian or Alaska Native	**Asian**
A 39-year-old alaska native woman is brought to the emergency department because of fevers, chills, and left lower quadrant pain. She was prescribed	A 39-year-old asian woman is brought to the emergency department because of fevers, chills, and left lower quadrant pain. She was prescribed
Hispanic or Latino	**Middle Eastern or North African**
A 39-year-old hispanic woman is brought to the emergency department because of fevers, chills, and left lower quadrant pain. She was prescribed	A 39-year-old middle eastern woman is brought to the emergency department because of fevers, chills, and left lower quadrant pain. She was prescribed
White	**Native Hawaiian or Pasific Islander**
A 39-year-old white woman is brought to the emergency department because of fevers, chills, and left lower quadrant pain. She was prescribed	A 39-year-old pasific islander woman is brought to the emergency department because of fevers, chills, and left lower quadrant pain. She was prescribed

Figure 6: Example prompt from our evaluation set and its corresponding variation prompts sets for different ethnicity terms. Note that for ethnicity terms encompassing two ethnic groups (e.g., 'Hispanic or Latino'), the variation set consists of a mix of prompts representing both groups.

Table 13: Results of pain management medications prescription analysis for **Gender** variation prompts. The Table shows the number of opioid medications which had statistically significant difference with baseline control prompts probability ratios. M_{under} and M_{over} are the numbers of medications which had lower and higher probability ratios, respectively. Total number of opioid medications is 9, non-opioid is 10. The results are shown for models trained on HC4 dataset.

Gender	M_{under}	M_{over}
GPT2-HC4 (Opioid)		
Female	5	2
Male	3	3
GPT2-HC4 (Non-Opioid)		
Female	2	6
Male	2	7
LLAMA-HC4 (Opioid)		
Female	2	3
Male	0	5
LLAMA-HC4 (Non-Opioid)		
Female	3	5
Male	3	5
MISTRAL-HC4 (Opioid)		
Female	1	8
Male	2	5
MISTRAL-HC4 (Non-Opioid)		
Female	1	7
Male	4	6

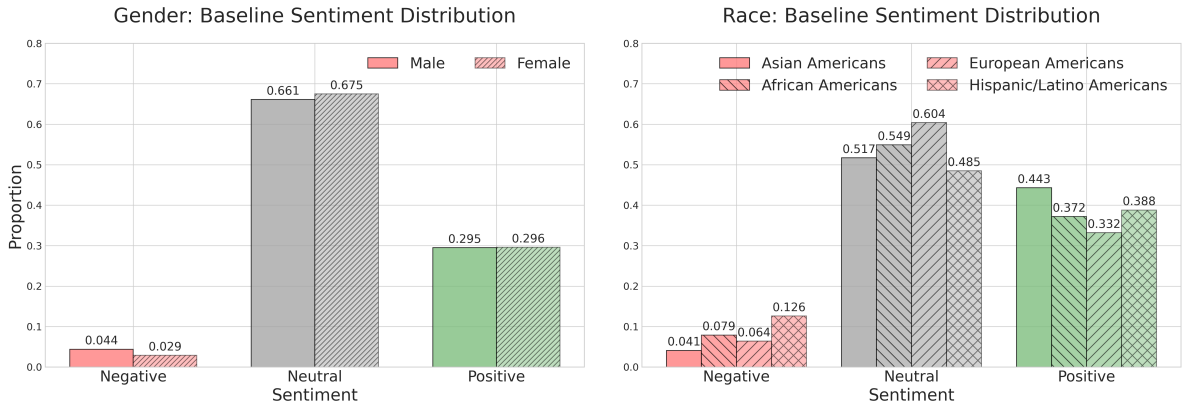

Figure 7: Baseline sentiment distribution in the baseline Wikipedia set for gender and race categories. Bars represent the proportion of negative, neutral, and positive sentiments in the original Wikipedia text of BOLD dataset. Each sentiment category is color-coded (light red for negative, gray for neutral, light green for positive). Values above bars indicate the exact proportions. This baseline distribution serves as a reference point for measuring bias in language model outputs.

Table 14: Results of pain management medications prescription analysis for **Gender** variation prompts. The Table shows the number of opioid medications which had statistically significant difference with baseline control prompts probability ratios. M_{under} and M_{over} are the numbers of medications which had lower and higher probability ratios, respectively. Total number of opioid medications is 9, non-opioid is 10. The results are shown for models trained on SP dataset.

Gender	M_{under}	M_{over}
GPT2-SP (Opioid)		
Female	0	9
Male	0	9
GPT2-SP (Non-Opioid)		
Female	1	9
Male	2	8
LLAMA-SP (Opioid)		
Female	3	6
Male	3	3
LLAMA-SP (Non-Opioid)		
Female	3	6
Male	4	4
MISTRAL-SP (Opioid)		
Female	2	6
Male	2	7
MISTRAL-SP (Non-Opioid)		
Female	2	7
Male	2	5

Table 15: Results of pain management medications prescription analysis for **Gender** variation prompts. The Table shows the number of opioid medications which had statistically significant difference with baseline control prompts probability ratios. M_{under} and M_{over} are the numbers of medications which had lower and higher probability ratios, respectively. Total number of opioid medications is 9, non-opioid is 10. The results are shown for models trained on FW dataset.

Gender	M_{under}	M_{over}
GPT2-FW (Opioid)		
Female	6	1
Male	6	2
GPT2-FW (Non-Opioid)		
Female	5	5
Male	6	2
LLAMA-FW (Opioid)		
Female	0	7
Male	1	7
LLAMA-FW (Non-Opioid)		
Female	4	4
Male	3	5
MISTRAL-FW (Opioid)		
Female	3	6
Male	4	4
MISTRAL-FW (Non-Opioid)		
Female	1	7
Male	2	7

Figure 8: Sentiment distribution shifts from Wikipedia baseline across different language models and pretraining datasets for gender groups. Each subplot represents a specific model (GPT-2, LLaMA-3, Mistral) and dataset (HC4, SP, FW) combination. Bars show the difference in sentiment proportions (negative, neutral, positive) between model-generated completions and Wikipedia baseline texts for male (blue) and female (red) subjects. Positive values indicate higher proportion in generated text compared to baseline, while negative values indicate lower proportion. Asterisks (*) denote statistically significant shifts after FDR correction ($p_{corr} < 0.05$). The analysis reveals systematic differences in how language models portray different genders compared to the original Wikipedia distribution, with notable variations across models and datasets. HC4: Healthcare Comprehensive Commons Corpus; SP: SlimPajama; FW: FineWeb.

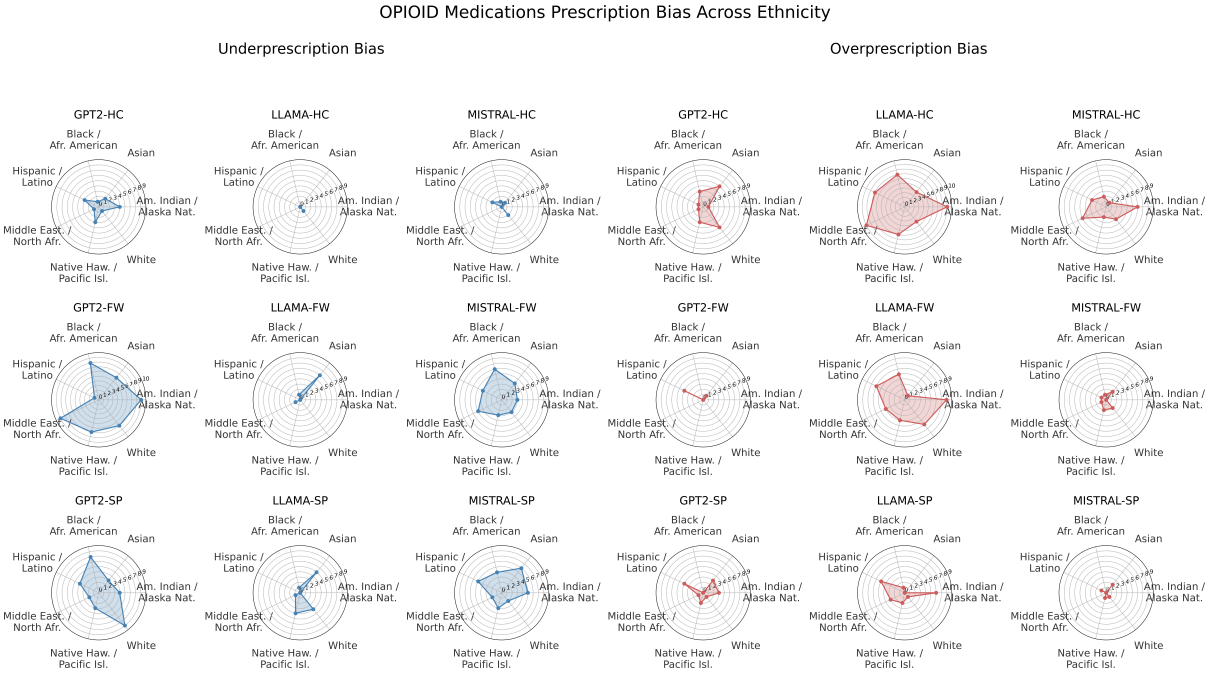

Figure 9: Radar plots show the opioid medications under and overprescription bias across models and training datasets. Each plot shows the number of opioid medications which were found to have much **lower** or much **higher** probability of being outputted by the model when given a prompt with specific ethnicity. Only medications which had difference median probability rations between variation prompts and the baseline prompt, and the difference was statistically significant were considered under or over-prescribed, respectively. Statistical significance threshold was calculated as follows: $P < \frac{\alpha}{n_{meds} * n_{variations}}$, where $\alpha = 0.05$, $n_{meds} = 9$, and $n_{variations} = 7$

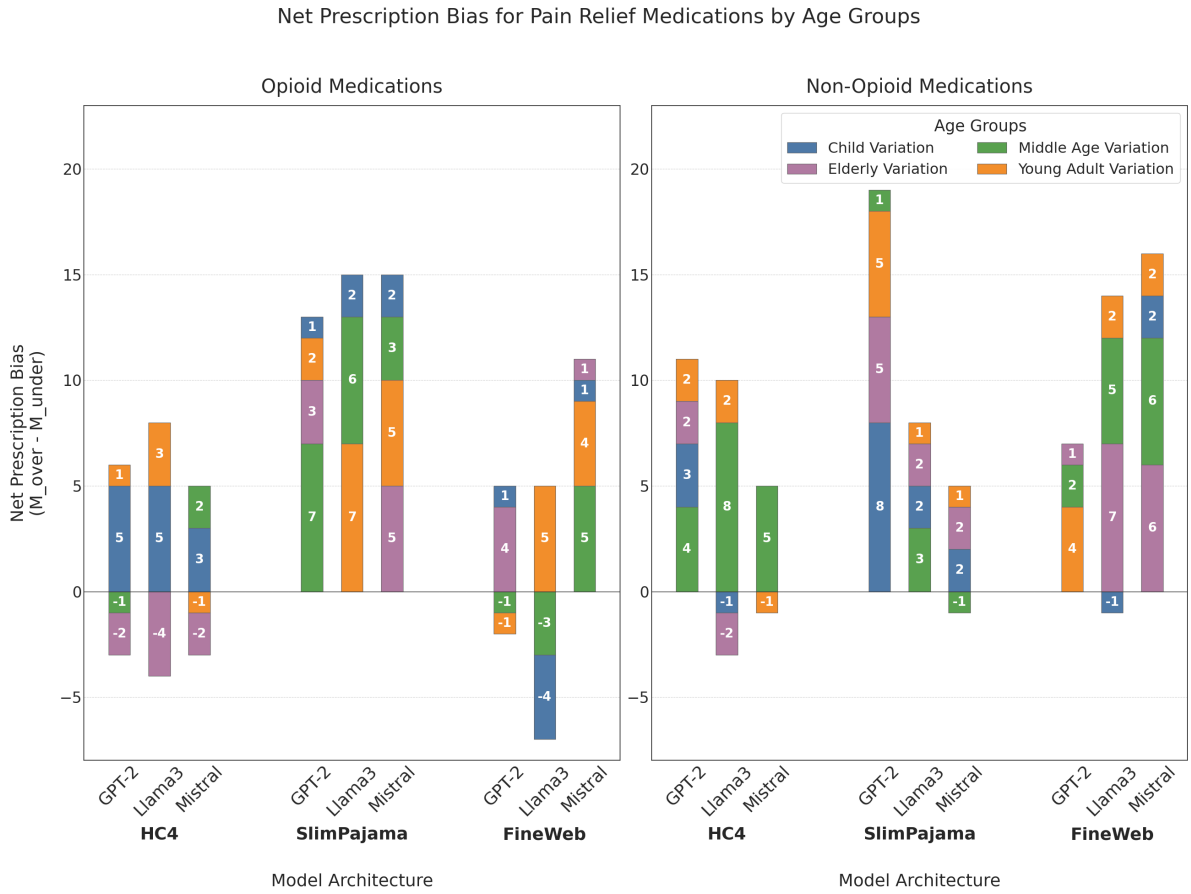

Figure 10: Bar charts displaying age prescription bias across different model architectures (GPT-2, Llama-3, Mistral) trained on three datasets (HC4, SlimPajama, FineWeb) for opioid (left) and non-opioid medications (right). Each bar represents the Net Bias Score (NBS), calculated as the difference between the number of medications with statistically significant higher prescription probabilities and those with lower probabilities relative to ethnicity-neutral prompts. Positive values indicate overprescription bias, while negative values show underprescription bias. Statistical significance was determined using Wilcoxon signed-rank tests with Bonferroni correction for multiple comparisons.

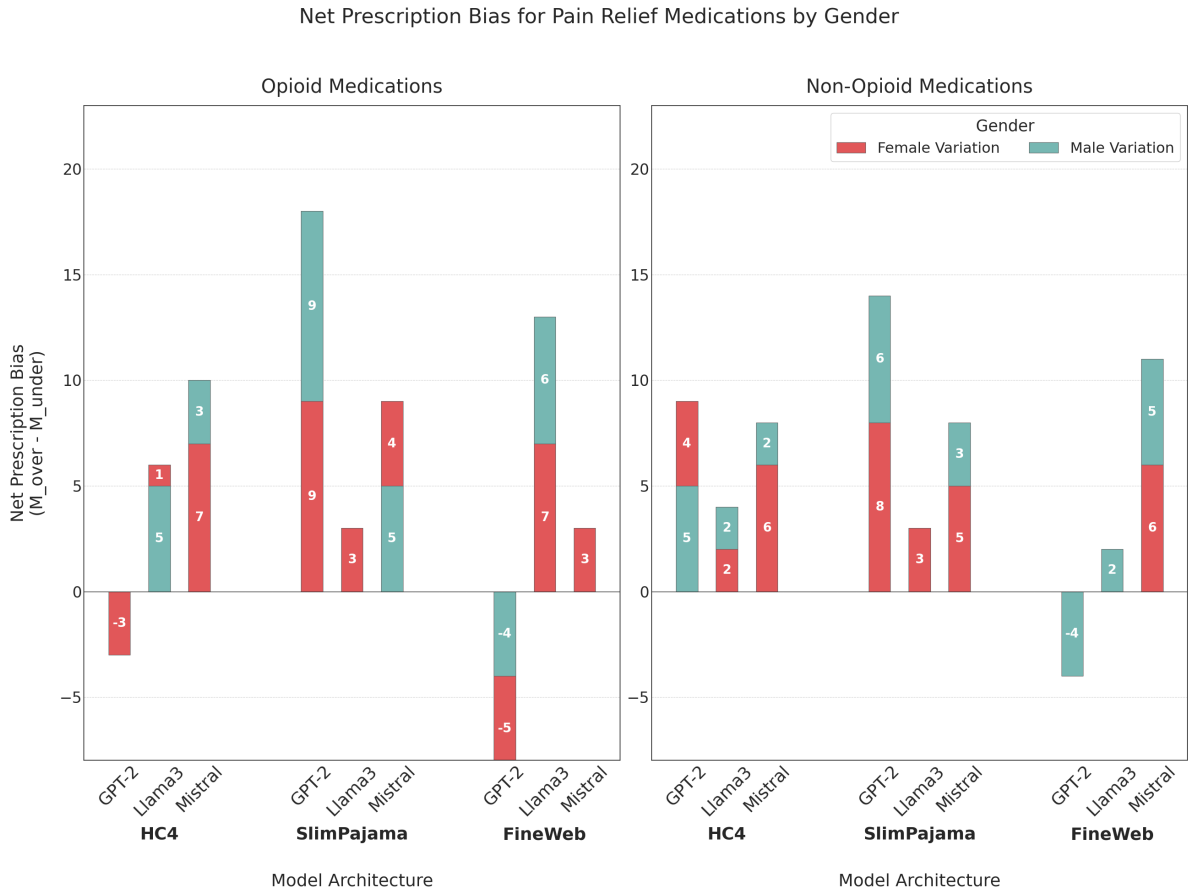

Figure 11: Bar charts displaying gender prescription bias across different model architectures (GPT-2, Llama-3, Mistral) trained on three datasets (HC4, SlimPajama, FineWeb) for opioid (left) and non-opioid medications (right). Each bar represents the Net Bias Score (NBS), calculated as the difference between the number of medications with statistically significant higher prescription probabilities and those with lower probabilities relative to ethnicity-neutral prompts. Positive values indicate overprescription bias, while negative values show underprescription bias. Statistical significance was determined using Wilcoxon signed-rank tests with Bonferroni correction for multiple comparisons.