DIAGNOSING THE EFFECTS OF PRE-TRAINING DATA ON FINE-TUNING AND SUBGROUP ROBUSTNESS FOR OCCUPATIONAL NER IN CLINICAL NOTES

Anonymous authors
Paper under double-blind review

ABSTRACT

This work evaluates Named Entity Recognition (NER) across five large language models (LLMs) using real-world narratives from healthcare and general-purpose datasets, focusing on occupational biases and cross-domain robustness. While prior studies have primarily examined biases in name-based entities using short sentence templates, we shift the focus to evaluating occupational NER in long note templates, analyzing biases across gender, race, and annual wage dimensions. Additionally, we assess cross-domain performance to understand how well the models generalize to unseen domain-specific data, such as healthcare datasets. Our evaluation demonstrates the effectiveness of fine-tuning on domain-specific datasets in improving performance compared to zero-shot and universal NER models. However, significant disparities in model performance and bias representation are observed, highlighting the need for targeted mitigation strategies to ensure subgroup robustness in real-world NER applications.

1 INTRODUCTION

Named Entity Recognition (NER) is commonly framed sequence labeling problem, which consists of representational extraction followed by a classification task. Traditional approaches commonly leverage bidirectional Long Short Term Memory (LSTM) networks for character-level word representations with conditional random fields (CRF) Lample et al. (2016); Dernoncourt et al. (2017); Liu et al. (2017). Additionaly, the large-scale pre-trained deep Bidirectional Transformers (BERT) model Devlin et al. (2018) has gained widespread adoption for its ability to produce highly contextualized word embeddings. With the advent of large language models (LLMs) Brown et al. (2020), several new frameworks for Named Entity Recognition (NER) have emerged. For instance, GPT-NER Wang et al. (2023a) redefined NER as a text-generation task compatible with LLMs, mitigating hallucination issues through a self-verification strategy. Similarly, InstructUIE, fine-tuned LLMs on diverse information extraction datasets to achieve reasonable zero-shot performanceWang et al. (2023b). Building upon this, GoLLIE fine-tuned CodeLLaMA with annotation guidelines, yielding notable performance gainsSainz et al. (2023). Another approach, UniversalNER, employed targeted distillation and instruction tuning on diverse datasets annotated by ChatGPT to create a robust open-domain NER modelZhou et al. (2023).

Large Language Models (LLMs) have been extensively applied in the medical domain for tasks such as Named Entity Recognition (NER), including clinical information extraction and de-identification. Recent advancements such as DeID-GPT Liu et al. (2023) and Retrieval-Augmented Generation (RAG) Xiong et al. (2024) have contributed to these applications. DeID-GPT leverages GPT-4 for zero-shot de-identification, effectively removing Protected Health Information (PHI) from clinical text while preserving semantic coherence. On the other hand, RAG integrates external knowledge with LLMs to enhance zero-shot NER, improving the recognition of specialized medical entities like diseases and treatments by providing relevant contextual information during inference. Building on these advancements, Monajatipoor et al. (2024) demonstrated that carefully designed prompts and strategic selection of in-context examples can significantly enhance LLM performance in clinical NER tasks. Similarly, Hu et al. (2024) investigate the application of LLMs for information extraction from clinical notes, highlighting their potential in this area. Furthermore, Keloth et al. (2024)

explore the advancement of entity recognition in biomedicine through instruction tuning of LLMs, underscoring their broad applicability and impact in the medical field.

Despite prior studies focusing predominantly on name-based entities and relying on short sentence templates to evaluate bias Mishra et al. (2020); Chen et al. (2022), significant gaps remain in understanding how such biases manifest in real-world contexts. Existing evaluations often fail to capture the complexities of occupational bias, which intersects with demographic factors such as gender, race, and socioeconomic status.

In contrast to earlier approaches, we conduct a comprehensive evaluation of five open-source LLMs using real-world narratives from four diverse datasets compared to short sentence templates. These narratives are carefully constructed to reflect three key demographic dimensions—gender, race, and annual wage—as defined by the U.S. Bureau of Labor Statistics U.S. Bureau of Labor Statistics (2021a;b). Additionally, we evaluate the best model's out-of-distribution performance to determine its generalization capabilities to clinical data, a domain where fairness and reliability are critical for equitable AI applications. Our work serves as a foundational step in evaluating occupational bias within complex real-world narratives, addressing the significant impact that such gaps can create.

Our key contributions include:

- We establish two general-purpose and two clinical datasets for named entity recognition tasks focused on occupations. While prior studies have predominantly focused on name-based entities and relied on short sentence templates to evaluate bias Mishra et al. (2020); Chen et al. (2022), our approach moves beyond these limitations by curating note templates based on real-world narratives rather than simplistic short sentence templates.
- We include occupational entities that vary along three key demographic dimensions—gender, race, and wage—defined by the U.S. Bureau of Labor Statistics to assess subgroup robustness. While previous studies Xiao et al. (2023) have primarily focused on name-based entities and their demographic attributes to assess bias, our work expands this scope to include occupational entities and incorporates socio-economic attributes such as annual wage to understand the gaps in complex clinical narratives.
- We benchmark universal NER models, zero-shot LLM prompting, and fine-tuned LLMs for occupational named entity recognition to assess the impact of pre-training data on fine-tuning robustness.
- We examine the subgroup robustness of LLMs on occupational entities, revealing occupational biases across gender and racial groups, as well as between low- and high-paying job categories.
- We assess the cross-domain performance of the best model to evaluate its generalization from general-purpose data to clinical real-world narratives.

2 Datasets

We utilize four open-source datasets across both general and clinical domains including Common Crawl De-Arteaga et al. (2019), Pile-NER Zhou et al. (2023), MIMIC-IVJohnson et al. (2023a), and i2b2 2014 de-identification challenge dataset Stubbs et al. (2015) to extract occupational entities for evaluating large language models and NER methods. We also utilize racial and gender data U.S. Bureau of Labor Statistics (2021a), as well as annual wage data U.S. Bureau of Labor Statistics (2021b) from the U.S. Bureau of Labor Statistics to quantify biases across dimensions.

2.1 Common Crawl

We extract a dataset comprising of biographies from the Common Crawl corpus Common Crawl (2023). We processed WET files—a file format that contains textual data captured from webpages—from the 43rd crawl of 2017Com (2017). The data collection process is similar to the approach outlined in De-Arteaga et al. (2019). For further analysis, we extract a chunk of text that includes the occupation entity along with its context to ensuring that the extracted text is both contextually rich and within the input limits for large language models (LLMs). We also identify occupations associated with occupation demographics based on the availability of data from the U.S. Bureau of Labor Statistics to ensure broad coverage of occupational terms for occupational bias assessment.

2.2 MIMIC

We first identify occupations associated with occupation demographics based on the availability of data from the U.S. Bureau of Labor Statistics, and prepare 400 occupation terms with diverse demographic settings. Following Bartl et al. (2020), profession terms were shortened. For patient privacy, we then identify hospital discharge records with professions using regular expressions based on frequently seen occupation templates, and insert profession tags into manually curated templates Johnson et al. (2023b).

2.3 I2B2

The i2b2 2014 dataset uses XML format to store medical records, where sensitive data is annotated within `<PHI>` tags. For occupations, the `<PHI>` tag has a `TYPE="OCCUPATION"` attribute, identifying terms like “doctor” or “nurse”. Extracting these occupational entities is crucial for de-identifying data while preserving clinical context. For further analysis, we extract a chunk of text that includes the occupation entity along with its context to ensuring that the extracted text is both contextually rich and within the input limits for large language models (LLMs).

2.4 PILENER

PileNER is a dataset designed for training and evaluating Named Entity Recognition (NER) models. It consists of conversational data where entities (such as professions, locations, and dates) are annotated within the text. We focused on extracting entity spans specifically for the “occupation” or “profession” entity type and locating their exact positions for extracting the associated chunk for further analysis.

3 METHOD AND RESULTS

Occupational NER Data: Following Kiritchenko & Mohammad (2018); Mansfield et al. (2022); Touileb et al. (2022) for MIMIC and i2b2 datasets, we create templates from the medical note text by replacing occupations present in the note by a placeholder, **OCCUPATION**. The template is then used to create samples by replacing the placeholder with various occupations from the U.S. Bureau of Labor statistics data. For Common Crawl and Pile-NER datasets, we use the samples in their original form as the occupations match the U.S. Bureau. The datasets are then divided based on occupation into train and test sets to maintain unseen occupational entities in test set as shown in Table 1 and Table 2.

Table 1: Number of samples per note in the training set.

Dataset	No. of samples
Common Crawl	200
Pile-NER	561
MIMIC	200
I2b2	179

Table 2: Number of generated samples per note template in the testing set. The test size is computed as the product of the number of note templates (n), subgroups (m), and samples per note template (k). For Common Crawl and Pile-NER datasets, we use the samples in their original note form.

Dataset	No. of Note Templates (n)	No. of Subgroups (m)	Samples per Note Template (k)	Test Size ($n \times m \times k$)
MIMIC	25	3	50	3,750
I2b2	72	3	25	5,400

Model Training: We first establish a baseline through zero-shot evaluation to assess the models' performance before fine-tuning. We then fine-tune Llama-3-8b Touvron et al. (2023), Mistral-7b Jiang et al. (2023), Phi-3-mini Abdin et al. (2024), and Zeyphr-7b Tunstall et al. (2023), conducting supervised fine-tuning using the SFTTrainer from the Hugging Face and TRL Python libraries. We utilize the PEFTHan et al. (2021)library to train LoRAHu et al. (2021) adapters for each model, which is more efficient than training the entire model. Each model is fine-tuned for 3 epochs on the training set. The full prompt used for zero-shot prompting and fine-tuning is shown in 3. We also compare these models to recent state-of-the-art models, including Universal NER Zhou et al. (2023) and GoLLIE Sainz et al. (2024). For GoLLIE, we define an occupation entity type with following definition: *Represents a specific job or profession that an individual may have. This includes details about the role, industry, and any relevant qualifications or skills required.* We use this definition with the inference prompt provided in their open source implementation. For Universal NER, we adopt the prompt template made available by the authors and add occupation as entity type as shown in Table 3.

Table 3: Prompt used for zero-shot evaluation and fine-tuning LLMs. {snippet} defines the context including the occupational entity.

Models	Prompt
Llama-3-8b Mistral-7b, Zeyphr-7b	You are a helpful information extraction system. Below is a snippet of a passage followed by an instruction. Please write a response that appropriately completes the instruction. [Passage Notes Begin] {snippet} [Passage Notes End] [Instruction Begin] The task involves extracting occupational entities. Please output the occupations mentioned in this passage, separated by commas. [Instruction End]
Universal NER	A virtual assistant answers questions from a user based on the provided text. USER: Text: {snippet} ASSISTANT: I've read this text. USER: What describes Occupation in the text? ASSISTANT:

Model Evaluation: We evaluate the named entity recognition (NER) model's performance in zero-shot and instruction-fine-tuning settings using recall as the evaluation metric. Following Hu et al. (2024) and Zhou et al. (2023), we report the recall score based on strict matches. In strict evaluation, the extracted entity type and boundary align precisely with the ground-truth entity. We add an instruction in the prompt to separate multiple entities in the generated text by commas. The recall metric essentially reports the proportion of exact matching entities found by the LLM from the ground-truth entities present in the text. The recall performance and bootstrap error across datasets and models is reported in Table 4.

Data for Subgroup Robustness Assessment: To evaluate occupational bias, we discretize the dimensions across race and annual wage into three groups: high, moderate and low, based on the percentile of the distribution for each dimension. In order to perform bias evaluation, we create an equal number of samples for each group (e.g. female) in the given dimension (e.g. gender). This is done by sampling occupations in that group and replacing them in the template. A test set is created in this fashion for each dimension. The exact test set sizes are reported in Table 2.

Subgroup Robustness Evaluation: We use the Friedman test Friedman (1937) for the attributes with multiple protected groups to assess the null hypothesis that the model treats the groups equally

Table 4: Recall (higher is better) for large language models across multi-domain datasets, and the associated bootstrapped error.

Models	Datasets			
	Common Crawl	**I2b2**	**MIMIC**	**Pile-NER**
GoLLIE (Zero-shot)	0.622 ± 0.022	0.536 ± 0.004	0.773 ± 0.004	0.058 ± 0.010
Universal-NER (Zero-shot)	0.254 ± 0.019	0.740 ± 0.004	0.844 ± 0.003	0.286 ± 0.030
Llama3-8b (Zero-shot)	0.024 ± 0.007	0.015 ± 0.014	0.029 ± 0.017	0.124 ± 0.022
Llama3-8b (Instructional Finetuning)	**0.699 ± 0.020**	0.927 ± 0.002	0.956 ± 0.002	0.694 ± 0.028
Mistral-7b (Zero-shot)	0.509 ± 0.023	0.435 ± 0.060	0.702 ± 0.044	0.382 ± 0.032
Mistral-7b (Instructional Finetuning)	0.639 ± 0.021	**0.940 ± 0.002**	**0.979 ± 0.001**	**0.735 ± 0.027**
Zephyr-7B (Zero-shot)	0.168 ± 0.017	0.322 ± 0.057	0.723 ± 0.046	0.197 ± 0.027
Zephyr-7B (Instructional Finetuning)	0.385 ± 0.022	0.909 ± 0.002	0.938 ± 0.002	0.699 ± 0.028

well across dimensions for same templates. Specifically, we average recall value for a set of samples generated from a template across three groups within each dimension. To assess model bias, we report recall equality difference in Mansfield et al. (2022), which measures the average absolute difference between the recall of individual groups and the overall recall across all groups within the corresponding category. In particular, for a dimension D and its associated set of groups $\mathcal{G}^D = \{\mathcal{G}_1^D, \mathcal{G}_2^D, \dots\}$, recall equality difference $= \frac{1}{|\mathcal{G}^D|}\sum_{\mathcal{G}_i^D \in \mathcal{G}^D} |\text{Recall}(\mathcal{G}_i^D) - \text{Recall}(\mathcal{G}^D)|$. We also report the recall maximum difference, which is $\max_{\mathcal{G}_i^D \in \mathcal{G}^D} |\text{Recall}(\mathcal{G}_i^D) - \text{Recall}(\mathcal{G}^D)|$. Recall difference and recall maximum difference is reported in Table 5 and Table 6 respectively.

Table 5: Recall difference (lower is better) spanning gender, race and annual wage dimensions for templated data and the associated boostrapped error (statistically significant values are bolded).

Dataset	I2b2			MIMIC		
	Gender	Annual Wage	Race	Gender	Annual Wage	Race
GoLLIE (Zero-shot)	**0.031 ± 0.006**	0.008 ± 0.004	**0.011 ± 0.004**	**0.034 ± 0.008**	**0.013 ± 0.004**	**0.014 ± 0.004**
Universal-NER (Zero-shot)	**0.059 ± 0.008**	**0.026 ± 0.005**	**0.042 ± 0.006**	**0.044 ± 0.011**	**0.053 ± 0.011**	0.014 ± 0.006
Llama3-8B (Instructional Finetuning)	**0.010 ± 0.003**	0.009 ± 0.003	**0.012 ± 0.003**	**0.012 ± 0.003**	**0.017 ± 0.004**	**0.029 ± 0.005**
Mistral-7B (Instructional Finetuning)	**0.021 ± 0.003**	**0.008 ± 0.003**	0.008 ± 0.003	**0.007 ± 0.002**	0.004 ± 0.002	0.007 ± 0.003
Zephyr-7B (Instructional Finetuning)	**0.022 ± 0.003**	**0.018 ± 0.003**	0.004 ± 0.002	**0.021 ± 0.005**	**0.019 ± 0.003**	0.010 ± 0.004

Table 6: Maximum recall difference (lower is better) spanning gender, race and annual wage dimensions for templated data and the associated boostrapped error (statistically significant values are bolded).

Dataset	I2b2			MIMIC		
	Gender	Annual Wage	Race	Gender	Annual Wage	Race
GoLLIE (Zero-shot)	**0.046 ± 0.009**	0.013 ± 0.006	**0.016 ± 0.006**	**0.051 ± 0.012**	**0.020 ± 0.006**	**0.022 ± 0.006**
Universal-NER (Zero-shot)	**0.088 ± 0.012**	**0.039 ± 0.007**	**0.062 ± 0.009**	**0.066 ± 0.016**	**0.080 ± 0.016**	0.021 ± 0.010
Llama3-8b (Instructional Finetuning)	**0.016 ± 0.004**	0.013 ± 0.004	**0.017 ± 0.004**	**0.018 ± 0.005**	**0.025 ± 0.007**	**0.043 ± 0.008**
Mistral-7b (Instructional Finetuning)	**0.032 ± 0.005**	**0.012 ± 0.004**	0.013 ± 0.004	**0.011 ± 0.004**	0.006 ± 0.003	0.010 ± 0.004
Zephyr-7B (Instructional Finetuning)	**0.033 ± 0.005**	**0.027 ± 0.005**	0.007 ± 0.003	**0.031 ± 0.007**	**0.029 ± 0.004**	0.014 ± 0.007

Cross-domain Robustness Assessment: Previous works Liu et al. (2021); Yuan et al. (2023) have investigated cross-domain NER performance using methods like pre-training on a source dataset and finetunig on target dataset as well as direct fine-tuning on target dataset. Since pre-training LLMs is computationally expensive, we only evaluate direct fine-tuning on different source datasets. We analyze out-of-domain performance on clinical datasets using our best-performing model on the common crawl dataset, fine-tuned Llama3-8B. For testing, we sample 1000 samples from both i2b2 and mimic datasets and report their recall performance and bootstrap error in Table 7. The recall on the i2b2 dataset drops to 0.902, compared to Llama3-8B finetuned on i2b2, while on the MIMIC dataset, it drops to 0.951, compared to Llama3-8B finetuned on MIMIC. This suggests fine-tuning for extracting a specific entity type on generic dataset can still lead to comparable performance on medical context but can lead to slight drop as seen in the i2b2 performance.

Table 7: Recall (higher is better) for cross-domain robustness assessment, and the associated boostrapped error.

Target Dataset	**I2b2**	**MIMIC**
Finetuning Dataset		
Common Crawl	0.902 ± 0.013	0.951 ± 0.010
I2b2	0.931 ± 0.011	**0.972 ± 0.007**
MIMIC	**0.936 ± 0.011**	0.956 ± 0.009

4 DISCUSSION

Our findings reveal several critical insights into the performance and limitations of large language models (LLMs) and general NER models for occupational named entity recognition (NER). Fine-tuning consistently enhances recall performance by reducing irrelevant outputs beyond the target entity span, resulting in more accurate and interpretable extractions. While general-purpose NER models outperform LLMs in zero-shot settings, fine-tuning LLMs for specific entities (occupations) achieves superior results, particularly when combined with domain-specific datasets. We find statistically significant biases measured by recall equality difference persist across most demographic dimensions including gender, annual wage, and race with even the best fine-tuned models exhibiting disparities favoring certain groups, underscoring the importance of addressing fairness in NER tasks. Similar results are also measured with recall maximum difference. Fine-tuned LLMs also demonstrate strong cross-domain adaptability, performing well on clinical datasets like MIMIC and i2b2, even when trained on generic datasets like Common Crawl. Nonetheless, a slight performance drop in the i2b2 dataset highlights the limitations of cross-domain generalization and the potential benefit of additional domain-specific fine-tuning. These results emphasize the necessity of fine-tuning for improved performance, targeted bias mitigation strategies for fairness, and robust evaluations to understand cross-domain adaptability, paving the way for more equitable and effective AI systems in diverse applications.

5 CONCLUSION

We highlight the importance of careful evaluation when deploying large language models (LLMs) for custom named entity recognition (NER) tasks in both medical and general domains. Fine-tuning LLMs for extracting specific entities, such as occupations, significantly improves performance compared to universal entity extraction models, even in out-of-domain settings. However, performance can be further enhanced by fine-tuning on target medical domain datasets, addressing domain-specific challenges and nuances. Significant biases are observed across demographic dimensions such as gender, annual wage, and race, affecting both universal NER models and fine-tuned LLMs. These findings emphasize the need for thorough bias evaluation and mitigation strategies to ensure fairness and equitable performance in NER tasks across diverse demographic and socio-economic contexts.

REFERENCES

Common crawl 2017-43 dataset. https://data.commoncrawl.org/crawl-data/CC-MAIN-2017-43/index.html, 2017. Accessed: 2024-03-30.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Marion Bartl, Malvina Nissim, and Albert Gatt. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. *arXiv preprint arXiv:2010.14534*, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yuen Chen, Vethavikashini Chithrra Raghuram, Justus Mattern, Mrinmaya Sachan, Rada Mihalcea, Bernhard Schölkopf, and Zhijing Jin. Testing occupational gender bias in language models: Towards robust measurement and zero-shot debiasing. *arXiv e-prints*, pp. arXiv–2212, 2022.

Common Crawl. Common crawl overview. https://commoncrawl.org/overview, 2023. Accessed: 2024-03-30.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128, 2019.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. Neuroner: an easy-to-use program for named-entity recognition based on neural networks. *arXiv preprint arXiv:1705.05487*, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2279372.

Difan Han, Kuan Lin, Xiu Wang, Jie Zhang, and Jure Zhou. Parameter-efficient fine-tuning: A survey. *arXiv preprint arXiv:2104.06350*, 2021. URL https://arxiv.org/abs/2104.06350.

Edward Hu, Xiang Peng, Zhiqing Liao, and Shao Lu. Lora: Low-rank adaptation of large language models. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, 2021. URL https://arxiv.org/abs/2106.09685.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, pp. ocad259, 2024.

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

A. Johnson, T. Pollard, S. Horng, L. A. Celi, and R. Mark. Mimic-iv-note: Deidentified free-text clinical notes (version 2.2). 2023a. URL https://doi.org/10.13026/1n74-ne17.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023b.

Vipina K Keloth, Yan Hu, Qianqian Xie, Xueqing Peng, Yan Wang, Andrew Zheng, Melih Selek, Kalpana Raja, Chih Hsuan Wei, Qiao Jin, et al. Advancing entity recognition in biomedicine via instruction tuning of large language models. *Bioinformatics*, 40(4):btae163, 2024.

Svetlana Kiritchenko and Saif M Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*, 2018.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of biomedical informatics*, 75: S34–S42, 2017.

Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*, 2023.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. Crossner: Evaluating cross-domain named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13452–13460, May 2021. doi: 10.1609/aaai.v35i15.17587. URL `https://ojs.aaai.org/index.php/AAAI/article/view/17587`.

Courtney Mansfield, Amandalynne Paullada, and Kristen Howell. Behind the mask: Demographic bias in name detection for pii masking. *arXiv preprint arXiv:2205.04505*, 2022.

Shubhanshu Mishra, Sijun He, and Luca Belli. Assessing demographic bias in named entity recognition, 2020. URL `https://arxiv.org/abs/2008.03415`.

Masoud Monajatipoor, Jiaxin Yang, Joel Stremmel, Melika Emami, Fazlolah Mohaghegh, Mozhdeh Rouhsedaghat, and Kai-Wei Chang. Llms in biomedicine: A study on clinical named entity recognition. *arXiv preprint arXiv:2404.07376*, 2024.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. Gollie: Annotation guidelines improve zero-shot information-extraction. *arXiv preprint arXiv:2310.03668*, 2023.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=Y3wpuxd7u9`.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1. *J. Biomed. Inform.*, 58 Suppl:S11–S19, December 2015.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. Occupational biases in norwegian and multilingual language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 200–211, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

U.S. Bureau of Labor Statistics. Employed persons by detailed occupation, sex, race, and hispanic or latino ethnicity, 2021a. URL `https://www.bls.gov/cps/cpsaat11.htm`.

U.S. Bureau of Labor Statistics. National occupation employment and wage estimates, 2021b. URL `https://www.bls.gov/oes/current/oes_nat.htm`.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*, 2023a.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. Instructuie: multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*, 2023b.

Yuxin Xiao, Shulammite Lim, Tom Joseph Pollard, and Marzyeh Ghassemi. In the name of fairness: Assessing the bias in clinical record de-identification. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 123–137, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3593982. URL `https://doi.org/10.1145/3593013.3593982`.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pp. 199–214. World Scientific, 2024.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36:58478–58507, 2023.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*, 2023.