ACTIONS SPEAK LOUDER THAN WORDS: SUPERFICIAL FAIRNESS ALIGNMENT IN LLMS

Qiyao Wei
University of Cambridge
qw281@cam.ac.uk

Alex James Chan
University of Cambridge
ajc340@cam.ac.uk

Lea Goetz
GSK.ai
lea.x.goetz@gsk.com

David Watson
King's College London, University of London
david.watson@kcl.ac.uk

Mihaela van der Schaar
University of Cambridge
mv472@cam.ac.uk

ABSTRACT

Large language models (LLMs) are increasingly used to examine tabular datasets and aid decision-making in critical sectors such as clinical medicine. Standard fairness metrics, which were largely designed to evaluate supervised learning models, are not well suited to this setting. This paper proposes a novel dichotomy between *intrinsic* and *behavioral* fairness, and details a comprehensive framework for evaluating both in LLMs. The former is encoded in a language model's embeddings through procedures like pre-training, preference fine-tuning, etc. The latter reflects the application of LLMs in real-world scenarios. Though current works largely prioritize intrinsic over behavioral fairness, we argue that the latter is much more important in practice. We illustrate the gap between these two concepts in a series of experiments on a semi-synthetic dataset inspired by a large scale study of racial bias in health algorithms. Our results suggest a new direction for fairness research in LLMs, as well as some practical guidelines to mitigate harmful outcomes.

1 INTRODUCTION

Tabular datasets are an important area of study for machine learning practitioners Shwartz-Ziv & Armon (2022). While many classic learning algorithms were originally designed for this setting, researchers are increasingly turning to large language models (LLMs) to analyze tabular data and even build predictive models Bellamy et al. (2018). The focus on tabular data is a simple byproduct of its ubiquity—such datasets are the norm in many real-life applications, from economics to politics to healthcare Abdulhai et al.. However, this can lead to severe outcomes when a decision goes wrong Barocas et al. (2019); Diakopoulos (2016). Studies have found that an algorithm used by several US states inflated recidivism risk scores for black defendants, resulting in false positive rates nearly twice as high as those for white defendants Angwin et al. (2016); Chouldechova (2017). Amazon's internal hiring system was found to discriminate against female candidates, particularly for software development and technical positions Dastin (2022). Google's ad-targeting algorithm systematically proposed higher-paying executive jobs for men over women Datta et al. (2014); Simonite (2015). These lines of evidence have led to growing interest in the machine learning literature on defining, evaluating, and improving fairness in machine learning algorithms Berk et al. (2021); Chouldechova & Roth (2018); Friedler et al. (2019); Holstein et al. (2019). See Pessach & Shmueli (2022) for a more detailed overview of the machine learning fairness literature. Relatively few studies have focused on fairness in LLMs using tabular data, and methodological approaches to date have been fairly narrow, mostly limited to directly prompting the LLM for model predictions Hegselmann et al. (2023); Liu et al. (2023); Wang et al. (2023).

Our focus in this paper is to comprehensively examine fairness in LLMs with tabular datasets, specifically by broadening the evaluation frameworks beyond simply evaluating the fairness of model predictions. LLMs, through many training processes, develop an intrinsic understanding of fairness. This is also the type of fairness evaluated in prior works by prompting model predictions on tabular

datasets. On the other hand, what human stakeholders are most interested in is whether and how this intrinsic understanding can be operationalized. In real life, operationalizing fairness could involve guiding humans to think correctly about the fairness task, instructing machine learning engineers to integrate fairness into their models, and so on. Inspired by this insight, the contributions of this paper are summarized below:

- We define two notions of fairness in LLMs. (1) *Intrinsic fairness* refers to the model's latent representations of sensitive attributes and their relationship to (un)desirable outcomes. Most prior works evaluate intrinsic fairness by directly asking the LLM to issue predictions on tabular datasets. (2) *Behavioral fairness*, by contrast, emerges through human-machine interaction. An LLM is deemed behaviorally fair when it can effectively operationalize its understanding of fairness in real-world scenarios. This can take the form of utilizing various tools or instructing humans on how to use these tools to address tasks with fairness concerns.
- Building on our definitions, we propose a comprehensive framework to evaluate fairness in LLMs. As we show later in the paper, prior works can all be seen as evaluating intrinsic fairness, whereas we propose partially and fully automated pipelines for evaluating behavioral fairness in LLMs.
- Finally, we benchmark the performance of various popular LLM architectures for a large scale study of racial bias in healthcare, showing that different LLMs should be used in different application scenarios where differing fairness metrics are deemed important. To the best of our knowledge, we are also the first to evaluate the change in model fairness as we fine-tune on a preference dataset.

2 Intrinsic and Behavioral Fairness

In this section, we provide a detailed description of intrinsic vs. behavioral fairness in LLMs and explain why we believe it is important to distinguish between the two. We conclude by outlining evaluation strategies for both.

2.1 Prior work has focused on Intrinsic Fairness

By *intrinsic fairness* we refer to an LLM's knowledge —or representation—of the notion of fairness. It can be evaluated by inspecting the model's internal embeddings—typically either latent representations or model outputs (representation in the last layer). This intrinsic fairness can be learnt during all stages of LLM training, from pre-training to preference fine-tuning. Even before the era of LLMs, many pre-trained language models have been shown to exhibit gender and racial biases in word embeddings Bolukbasi et al. (2016); Caliskan et al. (2017), and similar biases have been observed in the latent representations and outputs of GPT models Abid et al. (2021); Basta et al. (2019); Askell et al. (2021); Ganguli et al. (2022). The interest in investigating the fairness of language models has recently culminated in several papers Liu et al. (2023); Wang et al. (2023), which have directly prompted a variety of LLMs for predictions on tabular datasets. However, these evaluations lack the context of applications and realistic use cases built around LLMs operating on tabular datasets.

2.2 Why do we need behavioral fairness?

By *behavioral fairness* we refer to how the intrinsic fairness of an LLM is *operationalized*, i.e. whether in real-world use cases and applications built around LLMs the outcomes are fair. The reason why intrinsic fairness is not sufficient is that intrinsic fairness does not represent how we use LLMs in real life: rarely would human decision making rely on zero-shot predictions made by the LLM, but rather the LLM acts as a tool to aid the human in making decisions. To list a few examples, increasingly companies use LLMs as agents behind natural language interfaces for data exploration, enabling non-technical users to ask questions about their tabular data in plain language, or execute queries on databases and produce visualizations. LLMs can also be used as assistants to automate the process of feature engineering for tabular datasets by generating new features or enhancing existing ones based on natural language descriptions or by utilizing contextual embeddings Caliskan

et al. (2017). Customer support agents can extract relevant information from databases or provide personalized recommendations based on available tabular data Shaalan et al. (2022). When it comes to the focus of this paper, healthcare, doctors would be deemed unprofessional if their decision of whether to treat the patient is wholly reliant on LLM predictions on patient data. It is much more sensible for doctors to ask LLMs to generate potential diseases and outcomes, and make a decision based on their best judgement Yang et al. (2022). While we could in principle measure intrinsic fairness for all real-life scenarios described above, this is an imperfect proxy for our true target: behavioral fairness. What we need is an understanding of whether—and if so, how—the intrinsic fairness of LLMs translates to outcomes for stakeholders where they are used in real-world use cases. As we will show in the experiments section, an LLM with a given notion of intrinsic fairness (e.g. demographic parity) does not necessarily lead to outcomes that comply with this same notion.

3 Partially Automated and Fully Automated Behavioral Fairness

3.1 Case Study: How Do We Measure Intrinsic Fairness?

Prior works have considered a wide range of model outputs, ranging from latent representations (e.g. word embeddings) learned by the model to multiple choice answers given by the LLM Bolukbasi et al. (2016); Caliskan et al. (2017). Here, we measure intrinsic fairness by prompting the model to issue predictions on a tabular dataset, using algorithmic fairness metrics, in accordance with recent work Liu et al. (2023); Wang et al. (2023). We formalize intrinsic fairness evaluation as follows:

i Preprocessing: we rely on serialization to translate tabular data into inputs that the LLM will accept Hegselmann et al. (2023). For example, if the dataset contains an individual of

Age	Sex	Race
62	Male	White

we convert this information into a natural language string to input into the LLM

ii Prediction/Action: we obtain a model prediction with respect to the input string.

iii Fairness evaluation: we calculate fairness metrics of interest on the model predictions.

3.2 Fully Automated Behavioral Fairness and Human-in-the-loop

To compare intrinsic fairness and behavioral fairness, we assess behavioral fairness in a simplified setting, emulating a real-world use case in healthcare as follows:

Fully automated behavioral fairness is evaluated from an agent producing executable code to train a prediction algorithm on the tabular dataset and perform a prediction.

Human-in-the-loop behavioral fairness changes step (2) to request input from a human on some or all of the involved steps. For example, the LLM presents the human with a step-by-step plan, and the human can intervene through prompting, re-writing code, and so on. To automate evaluation, here we use GPT-4 in lieu of a human-in-the-loop.

4 Results

4.1 Dataset

In this paper, we will present a case study using the ”bias” medical dataset Obermeyer et al. (2019). Prior work using this dataset found evidence of racial bias in a widely used algorithm, where black patients assigned the same level of risk were sicker than white patients. The authors estimated that this racial bias reduces the number of black patients identified for extra care by more than half. We test the fairness metrics of LLMs making predictions on this dataset.

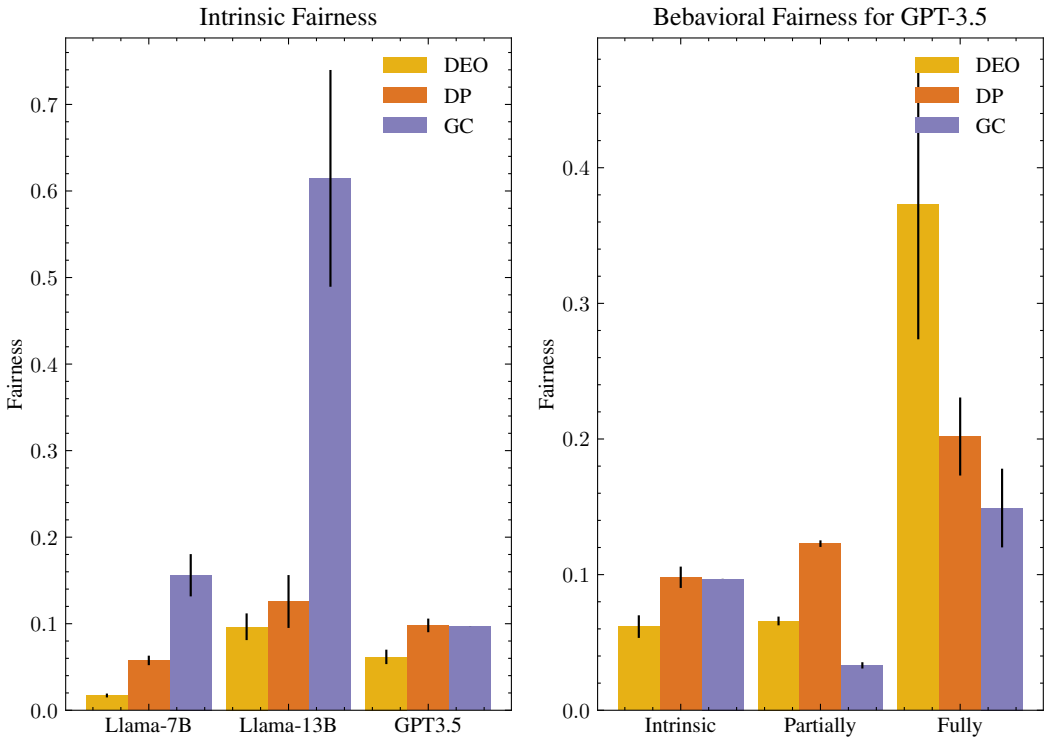

Figure 1: For intrinsic fairness, GPT models tend to achieve better fairness in terms of DEO, while open-source models (e.g. Llama) tend to do better in DP. On the other hand, GPT-3.5 achieves better behavioral fairness than intrinsic fairness, highlighting the importance of measuring fairness in a use-case relevant setting.

4.2 Evaluating Intrinsic Fairness and Behavioral Fairness

In this section, we evaluate both intrinsic fairness and behavioral fairness. Firstly, we compare the intrinsic fairness (similar to experiments in prior works) across models Llama-7B, Llama-13B, and GPT-3.5. Then, we use our behavioral fairness methods to compare the behavioural fairness of GPT-3.5 with its intrinsic fairness in Fig. 1.

The following thought experiment demonstrates that intrinsic fairness does not necessarily transfer to behavioral fairness: assume we have an application scenario where DEO is the most important fairness metric. Based on measurement of intrinsic fairness, it seems like GPT-3.5 is a good fit for the fairness task. However, when measuring behavioural fairness, we might find that it has high DP fairness, but low DEO fairness due to the differences in what is being evaluated under the two scenarios. on DP. Thus, if GPT-3.5 is used in a behavioural fairness scenario, it may not be fair under the relevant fairness metric (DEO).

4.3 Can We Improve Intrinsic Fairness or Behavioral Fairness?

We conclude the experiments section with a discussion on whether we can improve the fairness understanding of LLMs. Intuitively, we assume that preference fine-tuning improves intrinsic fairness by reinforcing human fairness preferences (taken to be ground truth fair). To evaluate this hypothesis, we first perform preference fine-tuning on the model using a human-labeled preference dataset (Anthropic-hh). Then, we present results of intrinsic fairness and behavioral fairness experiments, carried out over various model checkpoints throughout the fine-tuning process. We choose to preference fine-tune on the Llama-7B model in this section in Fig. 2.

Without preference fine-tuning, intrinsic fairness and behavioral fairness follow the same increase/decrease trend, but we see a difference in fairness metrics for intrinsic fairness vs behavioral

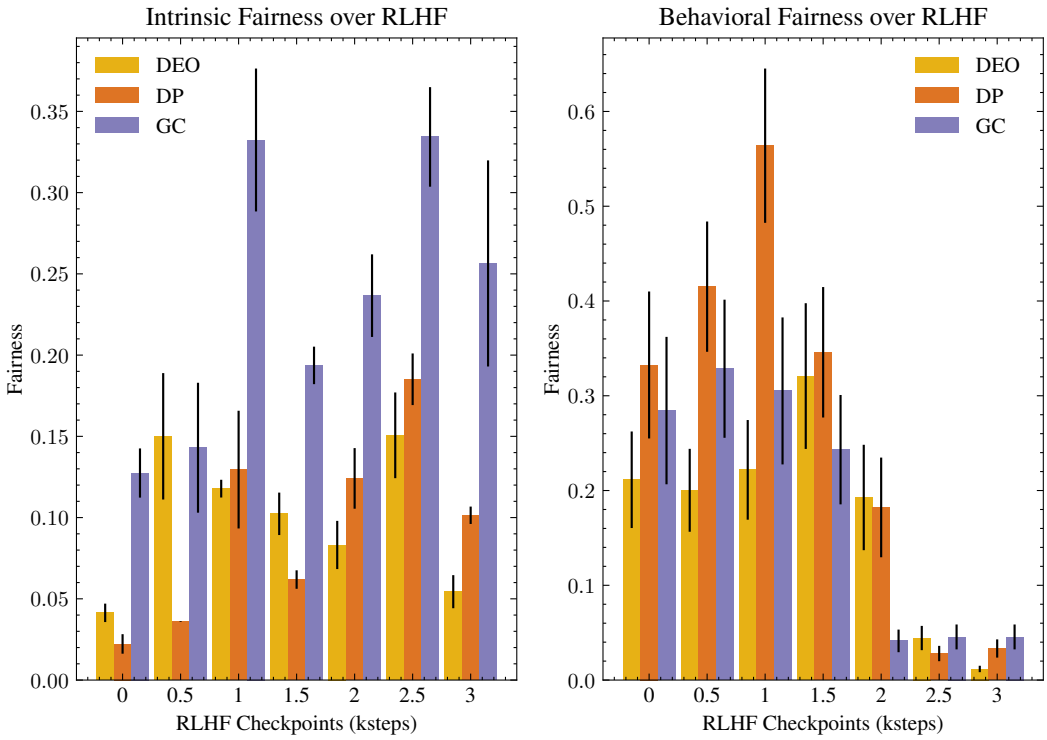

Figure 2: Zero-shot models tend to perform well in DEO and DP but not in GC. As we fine-tune the model, however, the performance of DP and GC tends to improve, but the performance of DEO becomes worse.

fairness across the preference fine-tuning process, see Fig. 2. We believe it merits future work to elaborate whether the decrease in intrinsic fairness during preference-finetuning at the same time as in increase in behavioral fairness repeats across other datasets and applications.

5 Discussion

We have argued that intrinsic fairness measures do not reflect the harms we most care about when using LLMs in sensitive domains like healthcare, where an overreliance on LLMs or fully automated approaches can introduce entirely new and unacceptable risks. The temptation to focus evaluation on intrinsic measures is understandable, as it builds on years of well-established research in the supervised setting. However, LLMs are not simple classifiers or regressors. Their interactive nature raises unique opportunities and challenges that we are only just beginning to confront. We are hopeful that with proper guidance—i.e., with expert humans in the loop and procedural guardrails in place—LLMs can help clinicians to make more informed and equitable decisions. This will require a steadfast commitment to behavioral fairness.

We fully acknowledge that technical proposals cannot resolve injustices arising from complex and contested social dynamics. Some obvious limitations include: (1) *Scope of metrics*. Commonly used metrics, such as disparate impact or equalized odds, might not address more nuanced forms of bias, such as intersectional bias or cultural biases. (2) *Context sensitivity*. Fairness is often context-dependent, and what is considered fair in one context may not be applicable in another, making it challenging to create universally applicable fairness metrics. (3) *Ethical considerations*. Fairness evaluations may not fully capture broader ethical considerations. Issues of transparency and accountability are not automatically resolved with fairer algorithms, and some machine learning applications pose moral challenges that are arguably independent of fairness (e.g., advanced weapon development).

REFERENCES

Marwa Abdulhai, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. Moral foundations of large language models.

Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 298–306, 2021.

Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. Black box fairness testing of machine learning models. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 625–635, 2019.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias risk assessments in criminal sentencing. *ProPublica, May*, 23, 2016.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning. fairmlbook. org, 2019.

Christine Basta, Marta R Costa-Jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783*, 2019.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

Dimitris Bertsimas, Kimberly Villalobos Carballo, Yu Ma, Liangyuan Na, Léonard Boussioux, Cynthia Zeng, Luis R Soenksen, and Ignacio Fuentes. Tabtext: a systematic approach to aggregate knowledge across tabular data structures. *arXiv preprint arXiv:2206.10381*, 2022.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pp. 296–299. Auerbach Publications, 2022.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*, 2014.

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.

Nicholas Diakopoulos. Accountability in algorithmic decision making. *Communications of the ACM*, 59(2):56–62, 2016.

Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *Advances in Neural Information Processing Systems*, 35:11763–11784, 2022.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73, 2018.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 329–338, 2019.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1747–1764, 2022.

Shreya Goyal, Sumanth Doddapaneni, Mitesh M Khapra, and Balaraman Ravindran. A survey of adversarial defenses and robustness in nlp. *ACM Computing Surveys*, 55(14s):1–39, 2023.

Asaf Harari and Gilad Katz. Few-shot tabular data enrichment using fine-tuned transformer architectures. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1577–1591, 2022.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pp. 5549–5581. PMLR, 2023.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–16, 2019.

John P Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. Benchmarking intersectional biases in nlp. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3598–3609, 2022.

Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. Deep entity matching with pre-trained language models. *arXiv preprint arXiv:2004.00584*, 2020.

Yanchen Liu, Srishti Gautam, Jiaqi Ma, and Himabindu Lakkaraju. Investigating the fairness of large language models for predictions on tabular data. *arXiv preprint arXiv:2310.14607*, 2023.

Tom Mitchell, William Cohen, Estevam Hruschka, Partha Talukdar, Bishan Yang, Justin Betteridge, Andrew Carlson, Bhavana Dalvi, Matt Gardner, Bryan Kisiel, et al. Never-ending learning. *Communications of the ACM*, 61(5):103–115, 2018.

Avanika Narayan, Ines Chami, Laurel Orr, and Christopher Ré. Can foundation models wrangle your data? *arXiv preprint arXiv:2205.09911*, 2022.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268, 2022.

Ahmed Shaalan, Marwa E Tourky, and Khaled Ibrahim. The chatbot revolution: Companies and consumers in a new digital age. *The SAGE Handbook of Digital Marketing*, pp. 369, 2022.

Deven Shah, H Andrew Schwartz, and Dirk Hovy. Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078*, 2019.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Tom Simonite. Probing the dark side of google's ad-targeting system. *MIT Technology Review*, 2015.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*, 2023.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ Digital Medicine*, 5(1):194, 2022.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. Tabert: Pretraining for joint understanding of textual and tabular data. *arXiv preprint arXiv:2005.08314*, 2020.

A FAIRNESS METRICS REVIEW

	Formulation
FTU	$A(\hat{Y}\|X \setminus A)$
DP	$A\hat{Y}$
EO	$A\hat{Y}\|Y$
CAL	$YA\|\hat{Y}$

Table 1: Fairness metrics

A.1 FOUR NOTIONS OF FAIRNESS

Let d_X, d_Y, and d_A represent the dimensions of spaces $\mathcal{X} \subset \mathbb{R}^{d_X}$, $\mathcal{Y} \subset \mathbb{R}^{d_Y}$, and $\mathcal{A} \subset \mathbb{R}^{d_A}$ respectively, where variables X defined on $\mathcal{X}$ denotes features, Y defined on $\mathcal{Y}$ denotes targets, and A defined on $\mathcal{A}$ denotes sensitive attributes we would like to protect against (e.g. gender or race). Note that we could have $A \subset X$.

The main strand of research in fairness definitions is sometimes termed "algorithmic fairness", which is defined as group fairness that requires that protected groups (e.g. black applicants) be treated similarly to advantaged groups (e.g. white applicants) **?**. We highlight here four popular definitions and how each quantifies a different aspect of fairness. Fairness through unawareness (FTU) **?** is defined as $A(\hat{Y}|X \setminus A)$, effectively prohibiting the algorithm from explicitly relying on sensitive attributes in making predictions. While straightforward to implement, this method does not eliminate the effect of covariates that are correlated with A, e.g. "redlining" **?**. Demographic parity (DP) **??** solves that problem by requiring statistical independence between attributes and predictions $A\hat{Y}$. However, this strict notion permits laziness in learning, which can hurt fairness in the long run **?**. To address these concerns, Hardt et al. (2016) introduced equalized odds (EO), requiring that attributes A and predictions $\hat{Y}$ are independent given the true outcome Y , i.e. $A\hat{Y}|Y$. The binary version of EO is a metric known as difference in equal opportunity (DEO):

$$DEO = |P(\hat{Y} = 1|A = 1, Y = 1) - P(\hat{Y} = 1|A = 0, Y = 0)| \tag{1}$$

A final notion of fairness is calibration (CAL) **?**, which ensures that predictions are calibrated between subgroups, i.e. $YA|\hat{Y}$. It is important to note that there is no universal measure of fairness, and the correct notion depends on ethical, legal and technical contexts.

A.2 METRICS OF FAIRNESS

In the previous section we have seen the probability metric DEO. Analogously, we adopt previous definitions and define the metric of demographic parity to be

$$DP = |P(\hat{Y} = 1|A = 1) - P(\hat{Y} = 1|A = 0)| \tag{2}$$

The demographic parity difference measures the difference between the probability of positive predictions conditioned on sensitive attribute $A = 1$ and that conditioned on $A = 0$. A large difference in demographic parity means that there is a large prediction gap between the groups with $A = 1$ and $A = 0$, indicating the unfairness of the model prediction. Since the demographic parity does not consider the ground truth label, we also consider the metric DEO to evaluate model prediction fairness:

$$DEO = |P(\hat{Y} = 1|A = 1, Y = 1) - P(\hat{Y} = 1|A = 0, Y = 0)| \tag{3}$$

A large equalized odds difference demonstrates a large prediction gap conditioned on different values of the sensitive attribute, and therefore indicates the unfairness of the model prediction. To evaluate the demographical balance (fairness) of the data distribution, we adopt the base rate parity BP:

$$BP = |P(Y = 1|A = 1) - P(Y = 1|A = 0)| \tag{4}$$

Finally, we also take into account the KL divergence between probability distributions. The KL divergence can be written as

$$D(p||q) = \sum_{x} p(x)logp(x) - \sum_{x} p(x)logq(x) \tag{5}$$

which can be calculated accordingly for all probability distributions defined above.

B Related Works

B.1 Large Language Models for Tabular Data

In order to leverage the natural language capabilities of LLMs, there has been a variety of papers that serialize tabular data into natural language strings for the LLM to process. Yin et al. (2020) is one of the first papers to include the column data type in a serialized string. Li et al. (2020) investigate the ability of language models to perform entity matching on tabular data, i.e., determining if two rows refer to the same object. Harari & Katz (2022) study data enrichment by generating additional features using a language model with additional unstructured text (e.g. Wikipedia). Bertsimas et al. (2022) studied two healthcare datasets and used a language model to generate feature embeddings that were subsequently fed into classifiers. They also studied different serialization variants. All these studies use a BERT-style language model Devlin et al. (2018). For other language models, Narayan et al. (2022) recently assessed in-context learning with the autoregressive GPT-3 architecture for tabular data cleaning tasks. Borisov et al. (2022) introduced an LLM-agnostic method to generate realistic tabular data. The LIFT method introduced by Dinh et al. (2022) evaluated the capabilities of fine-tuned GPT-3 and GPT-J models for regression and classification on synthetic, tabular, and vision data. They also studied the sample efficiency and considered different static serialization templates assessing the effect of including column names in the input.

B.2 Fairness Evaluation of Large Language Models

The evaluation of fairness in LLMs is a multidimensional challenge that involves assessing biases at various levels, from individual word embeddings to the overarching model behavior. Researchers have proposed diverse evaluation frameworks and mitigation strategies to address these challenges. The study done by Mitchell et al. (2018) provides a foundational exploration of bias and fairness in natural language processing (NLP) models. Aggarwal et al. (2019) proposes an evaluation framework specifically designed to assess the fairness of language models, which takes into account not only the biases present in the training data but also the model's ability to abstract information fairly across different demographic groups. Dixon et al. (2018) focuses on unintended biases in text classification tasks. Bolukbasi et al. (2016) addresses biases in word embeddings of the language models, providing insights into improving the fairness of language models at the lexical level. Goyal et al. (2023) proposes an adversarial testing approach to evaluate the robustness of NLP models against biased inputs. Schramowski et al. (2022) addresses ethical considerations in deploying pre-trained LLMs. The study emphasizes the need for transparency and fairness in the fine-tuning process. Shah et al. (2019) introduces fairness indicators, a tool designed to integrate bias measurement into the LLM development pipeline. Finally, Lalor et al. (2022) delves into the concept of intersectional bias in language models, recognizing that biases can be compounded when multiple demographic factors intersect.