Large Language Models: A Case Study on Corporate Values and Transformative Technologies

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Sidhardh Burre

Spring 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor
Caitlin D. Wylie, Department of Engineering and Society

Large Language Models (LLMs) are machine learning (ML) models engineered to generate plausible digital assets, a mechanism so powerful that it has enabled LLMs to summarize, translate, and generate content in a human-like manner. However, within a week of the launch of Google's latest LLM offering, Gemini, users were able to prompt the model for historically inaccurate images: Nazis depicted as people of color (Robertson, 2024). Although Google would halt Gemini's image generation capabilities and issue a public apology, the damage was done. Once again, the opaque and unruly nature of LLMs had proven too much to handle, only this time for the largest tech company on the planet. More importantly, it highlighted a growing divide: while the tech industry hurtles towards widespread adoption of LLMs, the academic sphere emphasizes ethical considerations and further research to avoid LLM misalignment, revealing a deep schism in values and confidence concerning the future of LLMs.

As exemplified by the latest Google Gemini fiasco, LLMs have significant vulnerabilities: a propensity for reinforcing biases and an inability to adapt to rapid societal shifts—a deficiency laid bare by Google's high-profile error. To analyze this deficiency, I employ Value Sensitive Design (VSD) a technique that provides theories and methods to systematically "account for human values…throughout the design process" (Value Sensitive Design Lab, n.d.). This paper employs VSD as a theoretical lens to dissect the dual challenges of dataset bias and unresponsiveness to frame shift that are inherent to LLMs, delving into the roots of these issues and how they compare with historical technological precedents.

I have anchored this analysis in an exploration of seminal journal articles, news stories, and industry statements to construct a comprehensive view of the technical, social, and entrepreneurial ecosystems that shape LLM development. The central thesis posits that for LLMs to benefit society, their development must integrate ethical considerations, addressing LLM inclination towards perpetuating biases and latency in reflecting frame shift. By rectifying these shortcomings, LLMs will not only reflect human language but serve as tools for progress towards a just and equitable society.

II. **Data Selection Introduces Biases into LLMs**

Large Language Models (LLMs), such as OpenAI's GPT-3 and GPT-4, have demonstrated profound capabilities in interacting with databases and generating creative content, yet they inherit significant flaws from the biased and problematic distributions of internet-based data used in their training. Despite the amazing capabilities these models demonstrate, the underlying technology for these transformative products is deeply flawed. The sheer depth and breadth of data available over the internet has enabled companies to source the data to power data-hungry Machine Learning/Artificial Intelligence (AI) models. However, internet-based data has extremely problematic distributions. When models ingest Internet data as part of their training, they are prone to encode stereotypical and derogatory associations along minority statuses that are typically uncorrectable (Prabhu & Birhane, 2020; Fast et al., 2016). This encoding occurs because the internet users are often biased along axes of existing inequalities within society. In this section, I argue how dataset selection is a design decision, followed by an exploration of how this decision affects LLMs.

Despite the Internet's vastness, it lacks diversity in terms of representation, with certain groups being significantly over-represented. Users that are younger and from developed countries are highly involved on the internet and are thus overrepresented (Sidoti et al., 2024). This overrepresentation is particularly pronounced in vocal sub-sections of the internet. To illustrate this lack of diversity, let’s consider one of the largest forums on the internet: Reddit. A Pew Internet Research survey found that over 65% of Reddit users are men. Further, the ages of

18 to 29 compose over half of the Reddit population (Barthel et al., 2016). This suggests that developers that train models from Reddit data produce models that are likely to mirror male voices in their vocabulary, mannerisms, and more. A subsequent survey of Wikipedians found that only 9-15% of users are women or girls (Barera, 2020). This same study found that Wikipedia is over-representative of Western, English-speaking countries while neglecting Sub-Saharan Africa, and much of the Global South. When developers draw training data from such skewed sources, developers are doomed to produce similarly biased models.

The issue of Internet user bias becomes more significant when considering how human moderators govern platforms that feature user-generated content. Despite marketing efforts that present websites like Reddit, Wikipedia, and Twitter as open and accessible, key factors render them less accessible to marginalized groups. In an article from Dr. Leslie Kay Jones (2020), an assistant professor at Rutgers' Department of Sociology, she documents multiple cases of users receiving death threats on Twitter. Interestingly, moderators suspended the accounts receiving threats, while the accounts issuing threats continued to persist. Additionally, she found that "a wide range of overlapping groups including domestic abuse victims, sex workers, trans people, queer people, immigrants, medical patients, neurodivergent people, and visibly or vocally disabled people" (Jones, 2020) are subject to harassment on Twitter, creating a hostile environment for these groups. This moderation system results in a highly biased user base within Twitter, with the majority voices empowered and minority voices further marginalized. Because developers design models trained on skewed data for global use, the models often fail to represent diverse populations and can perpetuate stereotypes, further widening social divides.

On the developer side, LLM engineers take significant liberties with dataset filtering, resulting in the further marginalization of minority identities. Brown et al. (2020) trained the large language model GPT-3 on a subset of the Common Crawl dataset, specifically filtering for documents most similar to GPT-2's training data, primarily sourced from Reddit, Wikipedia, and some books. While this filtering was effective in removing content classified as “unintelligible”, it is unknown what else was removed from the dataset. The Colossal Clean Crawled Corpus, a popular dataset within the NLP field used in nearly 600 papers so far, has been cleaned by removing any data containing a word from a repository of about 400 “naughty”, “obscene” or “bad” words (*List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/en at master)*. While this list contains words related to racial slurs, white supremacy, and sex, it also features words used predominantly within the LGBTQ community such as “twink” and “bulldyke”. Therefore, while developers may filter out racially and sexually charged content using this list of words, the process may also exclude LGBTQ+ voices. This filtration robs the model of training data that reclaims slurs and describes marginalized sub-groups positively. When digested and parroted by LLMs, LLMs can reify these biases. In one study by Basta et al. (2019), LLMs were found to have skewed perceptions of the gender bias implicit in occupations. The LLM was found to classify “`receptionist’ and ‘librarian’ for female and ‘architect’ and ‘philosopher’ for male” (Basta et al., 2019, p. 5). While some may argue that these associations aren’t harmful outright, it can lead to LLMs generating misconstrued associations, thereby propagating and perpetuating existing gender biases.

Furthermore, the ramifications of LLM bias in medicine raise alarming implications of LLM failures in critical settings. Due to systemic barriers that decrease the usage of the medical system by people of Black and Hispanic/Latino descent, there is little medical data specific to these racial groups for LLMs to train on. In a peer-reviewed journal article, Zhang et al. (2020) created a baseline model to predict in-hospital mortality and patient phenotype prediction which they use to explore LLM bias on racial axes. The researchers prompted the model with different races to “generate medical context given patient race” (p. 2) such that the model generates the content for clinical notes after being prompted with race. The researchers find that “the modification of race generates a worse course of action for African American patients” (p. 2). They go on to postulate that “patients of Black and Hispanic/Latino descent suffer from lower recall than other groups” (p. 7), indicating that when dealing with these patients, the LLM is unable to perform as effectively due to “under-utilization of the healthcare system” (p. 2) by members of those races. When doctors, insurance companies, and hospitals rely on LLM output to make decisions, any mistakes made by these machines can lead to misdiagnoses and mistreatments, driving patients away and reinforcing existing trends.

Developers using biased data to train LLMs is a key design decision that is insensitive to the current values inherent within society. As seen in prior examples, LLMs are highly vulnerable to skews within their data, reflecting biases and perceptions encoded within. This makes dataset selection a key component of LLM design. VSD requires the incorporation of “human values into the design process” (Friedman & Hendry, 2019, p. 4) ensuring the alignment of the product’s values with user values. Such alignment requires developers to refine datasets to remove biases and skews or at least to acknowledge potential biases in LLM output. Through the framework of VSD, we can establish that there is a value incongruency between the design of LLMs and the values inherent in our society.

Industry leaders have heralded LLMs as the future of technology, but with these models causing and accelerating the reification of existing harmful social dynamics in everyday life, one must question how useful these tools are. The values inherent within LLM datasets and the values the LLMs subsequently reify are in direct contrast with the aspirational values our society holds. Without addressing how LLM technology can exacerbate social inequities, it is irresponsible to continue developing and distributing such systems without the realignment of the values core to LLM design.

III. LLM Unresponsiveness to Social Currents

Biased datasets are evident sources of ethical lapses at the inception of LLMs, but perhaps more insidiously, ethical breaches also stem from the ongoing maintenance of LLMs, which leads to their inability to adapt to evolving social dynamics. Language serves as a pivotal tool for social movements, shaping narratives that can galvanize public support and sustain momentum. However, the efficacy of language becomes compromised when developers use outdated datasets that do not reflect the evolving landscape of language and societal issues to train LLMs. Consequently, LLMs may reinforce the status quo, undermining the dynamics of social change with language that remains anchored in past contexts instead of aligned with present reality. To explore this point, I will analyze a sociological report, followed by internet coverage of the Black Lives Matter (BLM) movement. Further, I will explore the potentially harmful effects caused by LLMs' incapability to follow shifting frames.

In a paper published in the Journal of Qualitative Sociology, Francesca Polletta (1998) discusses how narratives shape social movements. In one example, she discusses the Knights of Labor, a union group that called for an inclusive labor movement. As part of their movement, the Knights of Labor led a failed strike that ultimately resulted in the demise of their movement due to a loss of momentum. Polletta (1998) argues that efforts to frame the strike as a portion of an ongoing movement would have been critical in generating a “sustaining narrative that could have made of the Knights’ setbacks an episode in a longer story of overcoming” (1998, p. 431). The Knights’ inability to sway popular sentiment generated a “narrative failure [that] played an important role in the movement’s demise” (p. 432). This example illustrates how changing narratives can be key to counter-culture movements. Without the ability to contextualize failings and successes within dynamic narratives, via framing, counter-culture movements can appear as isolated events instead of components of a larger movement.

In the same vein, a paper by Twyman et al. (2017) highlights how social movements work to directly shape the framings of the narratives that power movements. A "frame shift" refers to a change in the way events or issues are perceived and discussed within the public sphere, often because of deliberate efforts by social movements to influence public perception and dialogue. While this frame shift is evident live on the internet, it remains unrepresented within the training data used by LLMs, therefore, not mirrored in live LLM responses. Twyman et al. (2017) find that in the wake of the BLM movement, “Wikipedia reduced its average response time from 361 to 51 days” (p. 6), implying that Wikipedians reduced the latency between incident occurrence and article creation thus resulting in a pattern of intensified documentation. Further, these Wikipedia pages displayed intense discourse between editors on shaping articles to create the BLM narrative, specifically in regard to “situating the BLM article with respect to other related events” (p. 7). This indicates that the society’s view of the BLM movement was in flux, with members, opposition, and general society working to shape and reshape the narrative of the movement continuously. However, this rapid adaptation in narrative framing is not captured in the training data of LLMs, leaving them ill-equipped to respond to the nuances of contemporary social movements.

LLMs’ inability to adapt to shifting frames is exacerbated in situations where media coverage of a social movement is minimal or, even worse, focuses primarily on the extremes of a social movement. In a paper published by Professor McLeod (2007) of the University of

Wisconsin-Madison's School of Journalism and Mass Communication, he demonstrates how media coverage tends to bias towards identifying “potential threats and negative consequences of protests” (p. 3), specifically by creating “‘moral panics’ by exaggerating threats” (p. 3). During the anti-Vietnam War movement, McLeod (2007) argues that “media coverage emphasized violence, flag-burning, and counter-cultural elements of minority anarchist and anti-war protestors” (p. 3). While some may argue that this is merely the nature of journalism, when developers feed such intentionally polarized data to LLMs, LLMs misrepresent these events within their output. Using such biased data to train LLMs further entrenches outdated or skewed views, failing to present a balanced or current perspective on events.

As discussed above, to align technologies with human values, VSD directs us towards developing technologies that “co-evolve with social structure” (McMillan-Major et al., 2024). A possible roadblock to this goal of updating LLMs to reflect rapid shifts in societal narratives are the high computing costs associated with retraining. According to the CEO of OpenAI (the creators of Chat-GPT), the total cost of training GPT-4 (the most recent, powerful model) was over $100 million. This is a prohibitive cost considering how dynamic social currents are, necessitating nearly constant retraining. A potential solution might be adopting fine-tuning approaches that can update LLMs with current data without full retraining. Fine-tuning uses compute resources that are “~10% of the compute used for pre-training” (Davidson et al., 2023) indicating that fine-tuning is a cost-effective method of introducing new data to the model. While some may argue that the root problem of dataset bias is not solved through finetuning, I argue that since finetuning requires a smaller subset of data, it is easier to meticulously curate and adjust these smaller datasets to accurately capture the evolving nature of social discourses (Lyu et al., 2024). Therefore, finetuning approaches are a cost-effective method to keep LLMs up to date and address the problems described above. The lack of existing finetuning measures is not a failure inherent to LLM design, but instead, a product of current methods to maintain LLMs. From a VSD lens, these current methods have misaligned values ingrained in their design. Rectifying this issue requires the restructuring of values within LLM maintenance, specifically towards developing LLMs and LLM-adjacent processes that enable the perpetual modification of LLMs at a low cost.

The ability of LLMs to accurately reflect and contribute to social narratives is crucial to the understanding of social movements. The reliance on outdated training datasets not only undermines the potential of LLMs but also perpetuates existing societal biases, hindering progress and going directly against the aspirational values core to technological development. Therefore, it is essential to develop LLMs capable of keeping pace with the evolution of societal norms and narratives, ensuring that LLMs are a tool for progress instead of an anchor to the past.

IV. The Values that Shape LLMs

The influence of values in the development of LLMs extends beyond technical implications. Unlike the inherent properties of the technology, these values reflect the broader corporate environments shaping their creation. Comparing LLMs and Open Source Technologies (OSTs) yields significant differences in development approaches. By comparing the two, we can unveil the impact of foundational values on technological evolution.

Linux (developed by Ubuntu) is the prototypical open-source project, showcasing the substantial benefits of an approach centered around accessibility, modularity, and community-driven development. Celebrated for its ease of setup, stability, and low maintenance (Sharma, 2022), it was built as an alternative to Microsoft's paid OS at the time (MS-DOS). Ubuntu designed Linux to be an open-source, professional-grade OS that offered limitless customizability to a diverse array of consumers. Linux's design principle of modularity allows for extensive adaptation, catering to diverse needs – from streamlined versions like CentOS for cloud systems to specialized distro like Kali Linux for cybersecurity (Kumar, 2023). This flexibility highlights the core open-source values: accessibility, adaptability, and user empowerment.

The Ubuntu organization, the driving force behind Linux, is an Open-Source, non-profit community renowned for producing the user-friendly distro. Ubuntu keeps the software freely available, regularly updating it with security patches and software improvements. This commitment is rooted in the organization's mission, which emphasizes the freedom to manipulate the software at will, use of the software regardless of language or disability, and the licensing terms included with the software (Our Mission, n.d.). As a prototypical open-source project, Ubuntu's Linux showcases the substantial benefits of a development approach focused on accessibility, modularity, and community-driven innovation.

However, the landscape for LLMs (and AI more broadly) diverges significantly due to developers' capitalist motivations. Consider the company Anthropic, an AI start-up that has raised $7.3 billion in the last year. According to an article written by veteran reporters Griffith and Metz (2024), who cover venture capital and AI respectively, Anthropic's funding deal has "upended Silicon Valley's start-up deal-making." This hype-driven landscape has influenced the valuations of other AI start-ups as well, with OpenAI reaching a deal that values the company at 80 billion dollars, a "valuation tripled in less than 10 months" (Metz & Mickle, 2024). These billion-dollar evaluations are indicative of how much attention and resources society is directing towards LLM development.

Accompanying these billion-dollar price tags are investor expectations: expectations to deliver products faster than their competitors. In this winner-takes-all environment, where the first company to launch the next AI model maintains the majority market share, companies are willing to do anything at all to get in (and stay in) the game. This has played out most prominently with OpenAI. Because OpenAI was first to market, it has been able to retain nearly 40% of the total market share leaving the remaining competitors to duke it out for smaller pieces of the pie (The Leading Generative AI Companies, 2023). Notably, Microsoft, AWS, and Google have 30%, 8%, and 7% market share respectively. This high-pressure setting emphasizes the rapid development of LLMs, typically trained on biased datasets, and developer pressure for faster deployments as opposed to an equitable, tested deployments.

As previously mentioned, Google itself was a victim of investor pressure and was forced to release a half-baked product that was subject to a subsequent recall. Google's Gemini model recently created an "image of the US Founding Fathers which inaccurately included a black man" and when prompted to generate German soldiers from WWII, it "incorrectly featured a black man and Asian woman" (Kleinman, 2024). While the article recognizes how biased data is the cause of the problem, I argue that the problem is a deeper-rooted structural problem that skews the values involved in the development of AI.

The dichotomy between the academic pursuit of LLM safety and the industry's rapid development pace illustrates the stark misalignment of values. A paper by Microsoft researchers in 2011, that has received nearly 4000 citations, introduced formulations for defining algorithmic fairness. But since then, it has primarily been the work of academics, including papers such as "Algorithmic Decision Making and the Cost of Fairness" by Corbett-Davies et al. (2017), "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" by Bender et al. (2021), and “StereoSet: Measuring Stereotypical Bias in Pretrained Language Models” by Nadeem et al. (2021), that has pushed the envelope of LLM fairness research. And yet, OpenAI the company behind the largest publicly available GPT, has very little publicly available research dedicated to assessing LLMs in a fairness context (Research Index, n.d.). Researchers perform work on LLM fairness that is very immature compared to the technological advancement of LLMs. This indicates that our means far out-strip our understanding of the technologies at hand. Furthermore, academics primarily drive the major research effort to improve AI fairness, instead of the companies that possess the capital and means to develop these models in the first place. This disparity underscores a systemic issue: the prioritization of market dominance over ethical responsibility in AI development.

In contrast, the open-source ethos, exemplified by Linux and the Ubuntu organization, holds paramount the values of inclusivity, freedom of modification, and equitable access. This approach fosters an alignment of technological development with broader societal values. On the other hand, due to the incentivization of profit and market share of the first actor, AI startups face intense pressure from their corporate backers to develop, train, and launch their models as quickly as possible. In this race to train and launch models as quickly as possible, startups cut corners in data collection. When startups then train models on this data, biased along existing social divides, the resultant models are prone to parroting the biases within the data.

The divergent paths of LLMs and OSTs highlight a fundamental choice in technological development: prioritizing rapid commercial success and upholding principles of openness, fairness, and equity. Value Sensitive Design requires the development of tools that are “responsive to the needs of individuals, communities, fields, and society at large” (McMillan-Major et al., 2024, p. 14). As Linux’s success demonstrates, aligning technological development with open-source values can lead to widespread adoption, innovation, and a more equitable tech landscape, serving as a path towards aligning design methods to ensure that the final products not only push the boundaries of what is technologically possible but also promote a more inclusive and fair society.

V. Conclusion

This paper has critically examined two key failings of LLMs: inherent biases derived from training data and latency to societal shifts. It highlights the pressing need for a paradigm shift in the development of LLMs, advocating for models that not only understand human language but also embody the principles of equity and inclusivity. Current LLMs, as constructed by high-profile tech companies, replicate and magnify existing societal biases, thereby perpetuating inequalities rather than mitigating them. By leveraging Value Sensitive Design to compare LLMs constructed by high-profile tech companies and OSTs, I have shown that a more equitable approach to the development of such technology is possible in our market environment.

The significance of this research lies not only in its identification of fundamental flaws within the architecture and training of LLMs but also in its broader implications for the lack of alignment between academic research and industry progress. By addressing these issues, we can unlock new knowledge and solutions that ensure technological advancements contribute positively to societal progress. This includes the creation of more nuanced and adaptive LLMs capable of engaging with the complexities of human language and society in a manner that is fair and inclusive. Moreover, the integration of ethical considerations into LLM development opens the door to applications that can bridge social divides, foster global understanding, and empower marginalized communities.

Building on the findings of this study, several new research questions emerge, marking pathways for future inquiry. This includes determining how to integrate systematic, real-time data monitoring and feedback loops into LLM training processes to ensure continuous adaptation as well as how to operationalize the principles of VSD in the commercial contexts that dominate LLM development to reduce LLM vulnerabilities. By exploring these questions, future research can move us closer to the development of LLMs that are not only technologically advanced but also socially responsible and ethically grounded. This endeavor is not merely academic; it is a crucial step towards realizing the full potential of AI technologies to serve humanity's best interests, promoting a more equitable and understanding world.

References

Barera, M. (2020). Mind the Gap: Addressing Structural Equity and Inclusion on Wikipedia. https://rc.library.uta.edu/uta-ir/handle/10106/29572

Barthel, M., Stocking, G., Holcomb, J., & Mitchell, A. (2016, February 25). 1. Reddit news users more likely to be male, young and digital in their news preferences. *Pew Research Center's Journalism Project*. https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

Basta, C., Costa-jussà, M. R., & Casas, N. (2019). Evaluating the Underlying Gender Bias in Contextualized Word Embeddings (arXiv:1904.08783). arXiv. https://doi.org/10.48550/arXiv.1904.08783

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Benjamin, R. (2020). Race After Technology: Abolitionist Tools for the New Jim Code. Polity.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). *Algorithmic decision making and the cost of fairness*. https://doi.org/10.1145/3097983.309809

Davidson, T., Denain, J.-S., Villalobos, P., & Bas, G. (2023). AI capabilities can be significantly improved without expensive retraining (arXiv:2312.07413; Version 1). arXiv. http://arxiv.org/abs/2312.07413

Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Publishing Group.

Fast, E., Vachovsky, T., & Bernstein, M. S. (2016). *Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community* (arXiv:1603.08832). arXiv. https://doi.org/10.48550/arXiv.1603.08832

Friedman, B., & Hendry, D. F. (2019). Value sensitive design: Shaping technology with moral imagination. The MIT Press.

Griffith, E., & Metz, C. (2024, February 20). Inside the Funding Frenzy at Anthropic, One of A.I.'s Hottest Start-Ups. *The New York Times*. https://www.nytimes.com/2024/02/20/technology/anthropic-funding-ai.html

Sidoti, O., Gelles-Watnick, R., Faverio, M., Atske, S., Radde, K., & Park, E. (2024). Internet/Broadband Fact Sheet. *Pew Research Center: Internet, Science & Tech*. Retrieved January 29, 2024, from https://www.pewresearch.org/internet/fact-sheet/internet-broadband/

Jones, L. K. (2020, October 4). Twitter wants you to know that you're still SOL if you get a death threat—Unless you're…. *Medium*. https://medium.com/@agua.carbonica/twitter-wants-you-to-know-that-youre-still-sol-if-you-get-a-death-threat-unless-you-re-a5cce316b706

Kleinman, Z. *Why Google's "woke" AI problem won't be an easy fix*. (2024, February 28). https://www.bbc.com/news/technology-68412620

Kumar, R. (2023, February 15). *The Top 10 Linux Distros for Different Use Cases – TecAdmin*. https://tecadmin.net/top-linux-distros-for-different-use-cases/

List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/en at master · LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words. (n.d.). GitHub. Retrieved March 20, 2024, from https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en

Lyu, K., Zhao, H., Gu, X., Yu, D., Goyal, A., & Arora, S. (2024). *Keeping LLMs Aligned After Fine-tuning: The Crucial Role of Prompt Templates* (arXiv:2402.18540). arXiv. https://doi.org/10.48550/arXiv.2402.18540

McLeod, D. M. (2007). *News Coverage and Social Protest: How the Media's Protect Paradigm Exacerbates Social Conflict. 2007.*

McMillan-Major, A., Bender, E. M., & Friedman, B. (2024). Data statements: From technical concept to community practice. *ACM Journal on Responsible Computing*, *1*(1), 1–17. https://doi.org/10.1145/3594737

Metz, C., & Mickle, T. (2024, February 16). OpenAI Completes Deal That Values the Company at $80 Billion. *The New York Times*. https://www.nytimes.com/2024/02/16/technology/openai-artificial-intelligence-deal-valuation.html

Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 5356–5371). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.416

Our Mission. (n.d.). Ubuntu. Retrieved February 8, 2024, from https://ubuntu.com/community/ethos/mission

Polletta, F. (1998). Contending Stories: Narrative in Social Movements. *Qualitative Sociology*, *21*(4), 419–446. https://doi.org/10.1023/A:1023332410633

Prabhu, V. U., & Birhane, A. (2020). *Large image datasets: A pyrrhic win for computer vision?* (arXiv:2006.16923). arXiv. https://doi.org/10.48550/arXiv.2006.16923

Research index. (n.d.). Retrieved February 27, 2024, from https://openai.com/research

Robertson, A. (2024, February 21). *Google apologizes for "missing the mark" after Gemini generated racially diverse Nazis*. The Verge. https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical

Sharma, D. (2022, March 3). *8 Reasons Why Ubuntu Is the Ideal Distro for Linux Newcomers*. MUO. https://www.makeuseof.com/why-ubuntu-ideal-for-linux-newcomers/

The leading generative AI companies. (2023, December 14). IoT Analytics. https://iot-analytics.com/leading-generative-ai-companies/

Twyman, M., Keegan, B. C., & Shaw, A. (2017). Black Lives Matter in Wikipedia: Collaboration and Collective Memory around Online Social Movements. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1400–1412. https://doi.org/10.1145/2998181.2998232

Value Sensitive Design Lab. (n.d.). VSD Lab. Retrieved April 16, 2024, from https://vsdesign.org/

Zhang, H., Lu, A. X., Abdalla, M., McDermott, M., & Ghassemi, M. (2020). *Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings* (arXiv:2003.11515). arXiv. http://arxiv.org/abs/2003.11515

Zhang, J., Marino, C., Canale, N., Charrier, L., Lazzeri, G., Nardone, P., & Vieno, A. (2022). The Effect of Problematic Social Media Use on Happiness among Adolescents: The Mediating Role of Lifestyle Habits. International Journal of Environmental Research and Public Health, 19(5), 2576. https://doi.org/10.3390/ijerph19052576