BIDIRECTIONAL ALIGNMENT FOR INCLUSIVE NARRATIVE GENERATION

Ken Kawamura
Independent Scholar
ken_kawamura@alumni.brown.edu

ABSTRACT

Aligning Large Language Models (LLMs) for narrative generation demands more than model refinement. For narratives of marginalized communities, whose voices are historically silenced or distorted, a purely AI-centric alignment is insufficient. This tiny paper argues for bidirectional human-AI alignment, emphasizing critical human engagement alongside AI development. Through literary case studies—Virginia Woolf's Judith Shakespeare and Saidiya Hartman's Venus—we demonstrate that LLMs inherit and propagate historical biases, reflecting deep epistemic gaps. Addressing these requires human interpretation to recognize data limitations and embedded assumptions. True alignment for inclusive narratives necessitates both refined AI and informed human participation, fostering AI literacy and critical engagement with LLM outputs. This bidirectional approach is crucial for ensuring AI contributes meaningfully to representative storytelling, a key challenge for inclusive AI research.

1 INTRODUCTION

Aligning Large Language Models (LLMs) with human values, especially in narrative generation, is fundamentally challenging. "Aligned" narratives are deeply intertwined with the complexities of human experience, particularly for historically marginalized communities whose voices have often been silenced or distorted. This position paper argues that a unidirectional approach to AI alignment—solely focusing on imbuing AI with predefined human values—is insufficient. For fragmented or obscured values, common in historical and literary contexts, a more nuanced approach is required. In the spirit of the inclusive and accessible ICLR "Tiny Papers" track [1], this concise contribution offers a critical perspective.

Through literary case studies, we demonstrate the inherent limitations of solely optimizing AI to generate authentic narratives for marginalized voices. These examples underscore the necessity of **bidirectional alignment**: critical human engagement with LLM outputs is as crucial as refining AI models (Shen et al., 2024). This bidirectional approach, directly in line with the theme of the Bi-directional Human-AI Alignment Workshop [2], involves concrete actions: recognizing training data limitations, and interrogating embedded assumptions. As AI-generated narratives increasingly influence our understanding of history and culture, human agency in interpreting and contesting these narratives becomes paramount. Indeed, as this workshop emphasizes, true alignment fundamentally demands both technical AI advancements *and* informed human participation. This concise contribution seeks to advance this critical discussion, particularly towards more inclusive and human-centered research practices within the alignment community and beyond.

2 THE CHALLENGE OF ALIGNMENT: SILENCED VOICES

A fundamental challenge in LLM alignment is addressing incomplete or distorted representations of human values, especially for historically marginalized communities. LLMs, trained on existing textual data, inevitably inherit and propagate historical biases. For instance, less than 1 in 10 books

[1] https://iclr.cc/Conferences/2025/CallForTinyPapers
[2] https://bialign-workshop.github.io

in the Library of Congress from the 19th century were authored by women (Waldfogel, 2023). This is not merely a data quantity problem, but an epistemological one. Data absence often reflects deliberate silencing and misrepresentation, representing structural erasures rather than random gaps. "Filling the gaps" is thus insufficient. We illustrate this through Virginia Woolf's Judith Shakespeare and Saidiya Hartman's Venus, revealing profound difficulties in aligning LLM-generated narratives with suppressed historical realities.

2.1 VIRGINIA WOOLF'S *Shakespeare's Sister*

Virginia Woolf's *Shakespeare's Sister* in *A Room of One's Own* (Woolf, 1989) imagines Judith, William Shakespeare's equally talented fictional sister. Woolf argues Judith, lacking socio-economic means in a patriarchal society, would have been historically silenced, her genius unrecognized and unrecorded.

This exemplifies how structural inequities erase marginalized voices, rendering them historically invisible, including to LLMs. Trained on textual corpora, LLMs lack access to Judith's untold story, limiting their ability to authentically represent such perspectives. While stylistic mimicry (e.g., Shakespearean language (Jhamtani et al., 2017)) is possible, it fails to capture the lived experience of a woman in that era—an experience epistemically inaccessible to the model, absent from the historical record.

This thought experiment highlights a key limitation of a purely AI-centric alignment approach. Refining training data alone cannot address structural erasures. Judith's absence is not simply a dataset problem; it's an epistemic challenge demanding human interpretation. Fine-tuning LLMs stylistically cannot resolve this deeper issue of historical exclusion.

2.2 SAIDIYA HARTMAN'S *Venus in Two Acts*

Saidiya Hartman's *Venus in Two Acts* (Hartman, 2008) explores the erasure of enslaved women's narratives. Hartman reconstructs Venus's story, an enslaved African woman, facing epistemic constraints due to the near-total absence of records. Available accounts, authored by oppressors—slave traders and colonial officials—offer profoundly biased and incomplete portrayals.

Hartman's work exposes a critical challenge for both historians and AI alignment. When historical narratives are filtered through the oppressor's perspective, reconstructing marginalized voices risks perpetuating distortions. LLMs, tasked with "filling the gaps," necessarily draw from these skewed sources, potentially reinforcing historical erasures and violence.

Bidirectional alignment demands more than refining AI models; it requires fostering critical engagement with their outputs. Humans must discern when AI-generated narratives reflect epistemic gaps, not historical truths. Without this critical lens, AI risks reinforcing, not challenging, historical erasures.

3 CONCLUSION: THE NEED FOR BIDIRECTIONAL ALIGNMENT

The case studies of Judith Shakespeare and Venus underscore that aligning AI with human values for inclusive narrative generation transcends technical data augmentation or model refinement. It is fundamentally a human problem, demanding engagement with historical silences and biases. Bidirectional alignment is crucial to address this complexity.

Instead of solely focusing on aligning AI to predefined values, bidirectional alignment prioritizes equipping humans to critically engage with AI outputs. This necessitates developing AI literacy to recognize biases and epistemic gaps within LLM-generated narratives. True alignment is reciprocal: AI is refined to better reflect human values, while humans learn to interpret, contest, and enrich AI-generated content. This calls for research that actively bridges technical AI development with humanistic critical inquiry. Only this bidirectional approach enables AI to contribute meaningfully to truly inclusive and representative storytelling, and fosters a more equitable future for AI and society.

REFERENCES

Saidiya Hartman. Venus in two acts. *Small Axe*, 12:1 – 14, 2008. URL https://api.semanticscholar.org/CorpusID:144243349.

Harsh Jhamtani, Varun Gangal, Eduard H. Hovy, and Eric Nyberg. Shakespearizing modern language using copy-enriched sequence to sequence models. *ArXiv*, abs/1707.01161, 2017. URL https://api.semanticscholar.org/CorpusID:9737200.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions, 2024. URL https://arxiv.org/abs/2406.09264.

Joel Waldfogel. The welfare effect of gender-inclusive intellectual property creation: Evidence from books. Working Paper 30987, National Bureau of Economic Research, February 2023. URL http://www.nber.org/papers/w30987.

V. Woolf. *A Room of One's Own*. A Harvest book. Harcourt Brace Jovanovich, 1989. ISBN 9780156787338. URL https://books.google.com/books?id=1B-IkVg7MHAC.