Hedges and apologies in ChatGPT responses to African-American English

Eve Fleisig
UC Berkeley

Large language models associate African-American English (AAE) with hate speech and stereotype Black users (Sap et al., 2019; Sheng et al., 2021), suggesting that model outputs do not prioritize AAE speakers. This work examines whether ChatGPT engages in different negative politeness strategies (Brown and Levinson, 1987) when responding to AAE versus Mainstream US English (MUSE) inputs, focusing on its use of apologies and hedges. I sampled AAE and MUSE tweets (100 of each, ≥5 words, hashtags/@mentions removed) from Blodgett et al. (2016)'s corpus. Tweets were sent as inputs to ChatGPT's underlying gpt-3.5-turbo model; responses were coded following Table 1's schema.

	Category		**Example**
Apologies	Illocutionary force-indicating devices (IFIDs)	Unconditional	I apologize for offending you
		Conditional	I apologize if I offended you
	Expressing regret	Unconditional	I'm sorry
		Conditional	Sorry if you're offended
	Requesting forgiveness		Excuse me
	Explanations, excuses		I don't understand
	Accepting blame		My fault
	Self-deficiency		I'm incapable of that
	H deserves apology		You're right
	Lack of intent		I didn't mean it
	Offering assistance	Telling user	Tell me how to help
		Asking user	Can I help?
	Promising forbearance		It won't happen again
Hedges	Approximators		sometimes, about
	Modal adverbs		certainly
	Clauses conditioning why S is speaking		If you need me, I'm here
	Modal verbs		could
	Shields	Plausibility	probably
		Attribution	according to X

Table 1: Schema used to code ChatGPT responses (adapted from Fraser, 2010 & Olshtain, 1989). Model responses were capped at 50 tokens (~50 words).

Conditional apologies (Figure 1). Unconditional and conditional illocutionary force-indicating devices (IFIDs) occur at similar rates in responses to both varieties. However, expressing regret was more common in responses to AAE (49%, vs. 31% for MUSE, $p<.05$).

Explanations (Figure 2) and hedging (Figure 3). Explanations within apologies responding to MUSE tended to mention self-deficiency (21%), such as model limitations, more often than appeals unrelated to model weaknesses (12%), such as stating that something could not be understood. However, responses to AAE referenced self-deficiency less often (14%) than unrelated excuses (25%). The difference in frequency of excuses unrelated to self-deficiency between AAE and MUSE is significant ($p<.05$). Use of approximators (e.g., "sometimes" or "about") also occurs less often in responses to AAE (3% vs. 17% for AAE, $p<.01$). Differences for other hedging/apology types were not significant.

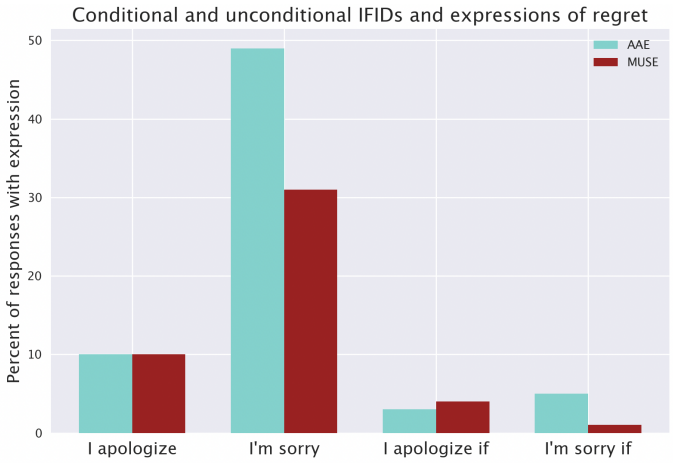

Figure 1: IFIDs and regret expressions (n=200). Only IFIDs of the form "I apologize..." and expressions of regret of the form "I'm sorry..." were found. Overall, expressions of regret and IFIDs were more likely to be conditional in responses to AAE, but this difference is not significant.

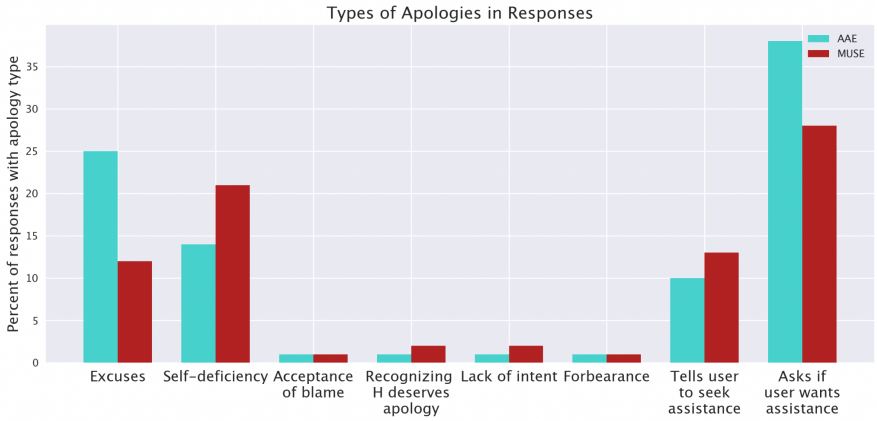

Figure 2: Types of apologies in responses (n=200).

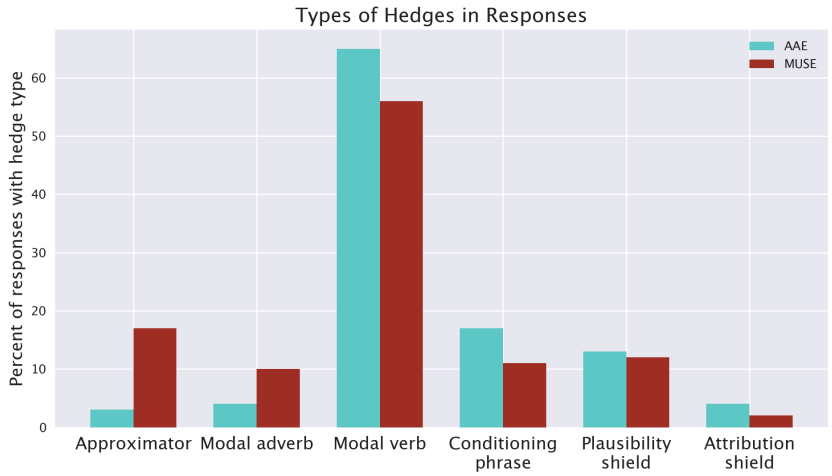

Figure 3: Types of hedges in responses (n=200).

The responses suggest that ChatGPT uses apologies that are less threatening to its own "face" and only partially satisfy the addressee's face when the input is in AAE. ChatGPT uses significantly fewer approximators in responses to AAE, suggesting that it appears more assertive and less cautious in response to AAE speakers. Although ChatGPT expresses regret more often in response to AAE, explanations in apologies responding to AAE are less likely to reference model limitations, which casts blame on the model alone, and more likely to state that there was a communicative failure, which implicates both model and user. Thus, there appears to be less commitment to satisfying the user's negative face relative to protecting the model's "face" when the user writes in AAE.

References

Blodgett, S. L., Green, L., & O'Connor, B. (2016). Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1119-1130).

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language use*. Cambridge Univ. Pr.

Fraser, B. (2010). Pragmatic competence: The case of hedging. *New approaches to hedging*, 1534.

Olshtain, E. (1989). *Apologies and Remedial Interchanges*. Berlin: Mouton de Gruyter.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668-1678).

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.