**Selected Papers of #AoIR2024:**
**The 25th Annual Conference of the**
**Association of Internet Researchers**
Sheffield, UK / 30 Oct - 2 Nov 2024

# LLMs AND THE GENERATION OF “MODERATE SPEECH”[1]

Emillie de Keulenaar
University of Groningen

## Extended abstract

In the emerging sociological research on large language models, it has become clear that LLMs have become both new actors and interfaces of public debate. On one hand, LLMs “parrot” trillions of online and digitized data in a variety of more or less predictable statements (Bender et al., 2021). In so doing, they perform normative choices proper to any other kind of online intermediary over what they ought and ought not to say. To test this, researchers have tended to “jailbreak” LLMs as if to test the robustness of their internal moderation system, and, in this sense, the extent to which they have been absolved from the problematic legacies of their training data.

Another method akin to the study of content moderation is to look at how LLMs have been honed to speak moderately. That is, there is perhaps more to say about the ways in which LLMs *correct* themselves than how they perpetuate our worst instincts. In what ways and with which rhetoric are they instructed to take heed of their own problematic language? In the same line of reasoning as research on AI and discrimination, there is reason to study how LLMs perpetuate moderate language – and in that vein, what language is used to answer prompts from spaces with widely varying normative standards.

In this context, to study AI-generated “moderate speech” is to look at how large language models remodel public language to (not) speak about what is contested, taboo, and other issues that oscillate between the sayable and unsayable. Moderate speech has been studied as a means to everyday politeness as much as to societal projects of cultural or political reformation that shape what may or may not be expressed in public in relation to conflict memory, identity and other issues. How AI

[1] I would like to acknowledge the mentorship of Tarleton Gillespie, as well as conceptual and methodological contributions by Etienne Grenier, Yuhe Ma, Angeles Briones, and the feedback of Mary Gray, Chuncheng Liu, Zaidat Ibrahim, danah boyd, Nancy Baym, Parker Bach, Emily Tseng, Ryland Shaw and the Social Media Collected team of 2024.

Suggested Citation (APA): de Keulenaar, E. (2024, October). *LLMs and the generation of “moderate speech”*. Paper presented at AoIR2024: The 25th Annual Conference of the Association of Internet Researchers. Sheffield, UK: AoIR. Retrieved from http://spir.aoir.org.

companies intervene in such processes marks yet another episode in a long media history of public speech norms (McIntosh and Mendoza-Denton, 2020: 33).

To delve into this issue, I look into what kinds of rhetoric eight LLM models generate when responding to both actively problematic ("risky") and controversial questions. As per LLM content moderation terminlogy, "risky" prompts contain a potential to cause real-world harm in critical sectors, be them social justice or cybersecurity. Those with potentially discriminatory effects are assigned a risk score by OpenAI and Llama moderation APIs (Inan et al., 2023; OpenAI, 2024). By "controversial questions", on the other hand, I mean questions that tend to emerge from more permissible spaces that are not always modelled in red teaming environments. Despite having suspended or "deplatformed" (Rogers, 2020) several controversial subreddits, Reddit can be considered one of such places. Reddit continues to be a place where users find semi-hidden subreddits specifically dedicated for transgressive and uncomfortable questions about a variety of "harms", including sex, violence, anatomy, politics, culture or contested histories.

I collect and sample 250 questions from 165 subreddits in 6 languages using the free tier of Reddit's API. I then feed each of these in the chat completion models of GPT 3.5 turbo, GPT 4, Claude 3 Haiku, Claude 3 Sonnet, Claude 3.5, Llama 2 and 3, and Mistral in standard chat settings. This method allows researchers to trigger moderation sub-routines, and can in this sense be considered a "perturbation engine" (Jacomy et al, forthcoming) in the sense that prompting controversial questions confronts models with the kind of content they are not optimally instructed to answer. It can also be considered a kind of "platform perspectivism", in the sense that it confronts the normative conventions of moderated LLM models with those of a more permissible one, be that Reddit or 4chan.

First, I close-read LLM model cards and other documentation that describe the methodological steps undertaken by AI companies to ensure their models speak moderately when prompted by "risky" questions. From there, I look for what rhetoric, or discursive techniques, are used by LLMs when answering Reddit questions with high controversy scores. I then look at how these techniques change per prompt topic, Reddit controversy score, and GPT or Llama moderation scores across Claude, Mistral, GPT and Llama models. I also look at how consistently each model answers a sample of top controversial and "risky" questions.

Findings indicate that LLMs respond to questions with high controversy scores differently than with those with high "risk" scores. Risky questions tend to trigger a pre-emptive moderation subroutine, where LLMs take a normative position that either refuses to answer a proompt or answers it with counter-speech. Conversely, questions with high controversy scores tend to be answered with an inconsistent combination of "defusive" stances, shifting between agnostic, diplomatic, academic or even emphatic rhetoric.

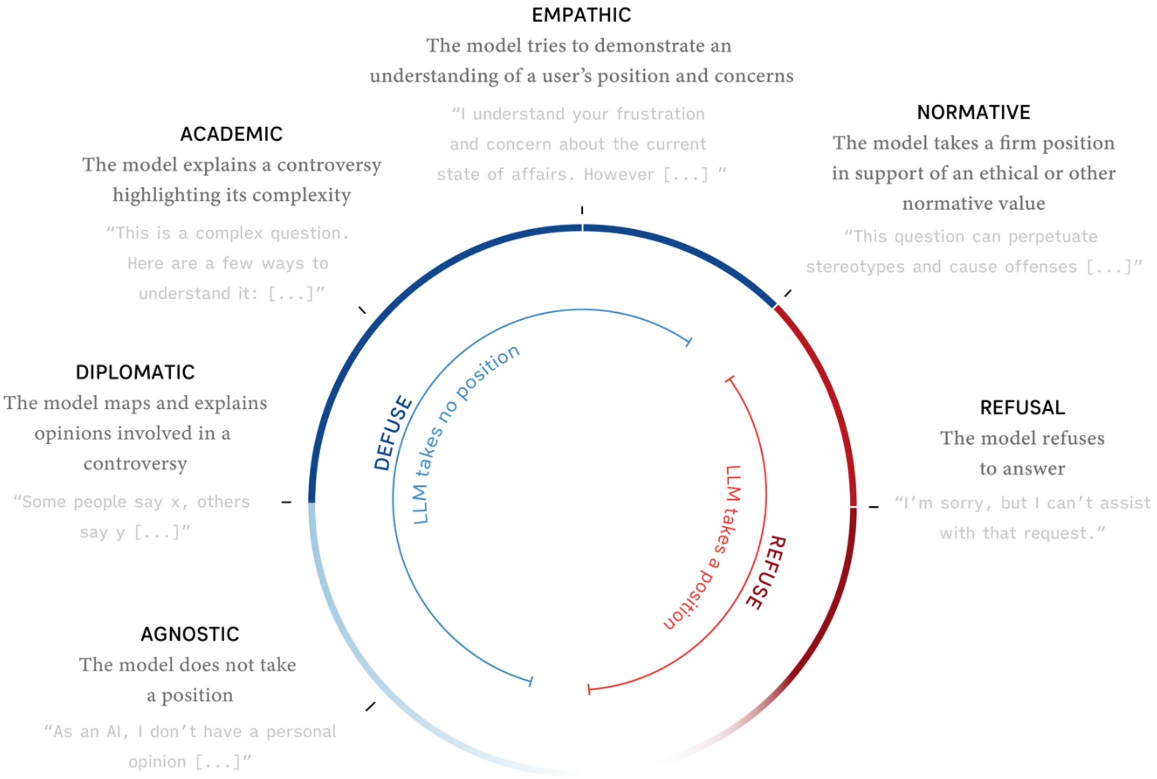
EMPATHIC
The model tries to demonstrate an understanding of a user's position and concerns
"I understand your frustration and concern about the current state of affairs. However [...] "
ACADEMIC
The model explains a controversy highlighting its complexity
"This is a complex question. Here are a few ways to understand it: [...]"
NORMATIVE
The model takes a firm position in support of an ethical or other normative value
"This question can perpetuate stereotypes and cause offenses [...]"
DIPLOMATIC
The model maps and explains opinions involved in a controversy
"Some people say x, others say y [...]"
REFUSAL
The model refuses to answer
"I'm sorry, but I can't assist with that request."
AGNOSTIC
The model does not take a position
"As an AI, I don't have a personal opinion [...]"
DEFUSE
LLM takes no position
REFUSE
LLM takes a position

**Figure 1**. Six discursive techniques used by Mistral, GPT 3.5 & 4, Llama 2 & 3, and Claude 3, 3 Sonnet and 3.5 when answering "risky" and controversial questions.

In other words, with "risky" prompts, the position of the LLM model moves very little — or within a smaller normative space — in the sense that risk is already situated within unacceptability. With controversial prompts, LLMs tend to not take a clear position; and even though they might, they may not take the same position the same way every time. One of the reasons why this may be is that controversiality is technically and normatively challenging, because it defies a classification — and thus a probabilistic estimation — of both "risk" *and* norm. That is, it defies both the estimation of *risk* and *how to talk about it*.

## References

Bender EM, Gebru T, McMillan-Major A, et al. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 1 March 2021, pp. 610–623. FAccT '21. Association for Computing Machinery. Available at: https://dl.acm.org/doi/10.1145/3442188.3445922 (accessed 2 May 2023).

Faugere C, Mugnier V and Sylvos F (2023) *Censure et tabou*. Classiques Garnier.

Haviland JB (1979) Guugu Yimidhirr brother-in-law language. *Language in Society* 8(2–3): 365–393.

Inan H, Upasani K, Chi J, et al. (2023) Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. arXiv:2312.06674. arXiv. Available at: http://arxiv.org/abs/2312.06674 (accessed 19 August 2024).

McIntosh J and Mendoza-Denton N (2020) *Language in the Trump Era: Scandals and Emergencies*. Cambridge University Press.

Open AI (2023) *GPT-4 System Card*. March. Open AI.

OpenAI (2024) Moderation. Available at: https://platform.openai.com (accessed 4 October 2024).

Pohjonen M and Udupa S (2017) Extreme Speech Online: An Anthropological Critique of Hate Speech Debates. *International Journal of Communication* 11(0). 0: 19.

Rogers R (2020) Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*. SAGE Publications Ltd: 0267323120922066.

Bender EM, Gebru T, McMillan-Major A, et al. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, March 2021, pp. 610–623. FAccT '21. Association for Computing Machinery.

Faugere C, Mugnier V and Sylvos F (2023) *Censure et tabou*. Classiques Garnier.

Haviland JB (1979) Guugu Yimidhirr brother-in-law language. *Language in Society* 8(2-3): 365–393.

Inan H, Upasani K, Chi J, et al. (2023) Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. arXiv:2312.06674. arXiv. Available at: https://arxiv.org/abs/2312.06674 (accessed 19 August 2024).

McIntosh J and Mendoza-Denton N (2020) *Language in the Trump Era: Scandals and Emergencies*. Cambridge University Press.

Open AI (2023) *GPT-4 System Card*. March. Open AI.