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

GPT4 V THE OVERSIGHT BOARD: USING LARGE LANGUAGE MODELS FOR CONTENT MODERATION

Nicolas Suzor
Queensland University of Technology

Lucinda Nelson
Queensland University of Technology

Introduction and background

Large-scale automated content moderation on major social media platforms continues to be highly controversial. Moderation and curation are central to the value propositions that platforms provide, but companies have struggled to convincingly demonstrate that their automated systems are fair and effective (Gillespie, 2018). In this paper, we set out to understand how the emergence of generative AI tools might transform industrial content moderation practices. We investigate whether the current generation of pre-trained foundation models may expand the established boundaries of the types of tasks that are considered amenable to automation in content moderation.

This paper presents the results of a pilot study into the potential use of GPT4 for content moderation. We use the hate speech decisions of Meta's Oversight Board as examples of covert hate speech and counterspeech that have proven difficult for existing automated tools. Our preliminary results suggest that, given a generic prompt and Meta's hate speech policies, GPT4 can approximate the decisions and accompanying explanations of the Oversight Board in almost all current cases. Our final paper will present analysis of several clear challenges and limitations, including particularly the sensitivity of variations in prompting, options for validating answers, and generalisability to examples with unseen content.

During the COVID-19 pandemic, in response to abrupt labour shortages and lockdown requirements, platforms accelerated their adoption of automated content moderation. These systems have primarily used machine learning classifiers trained on large datasets of human decisions and evaluated for consistency against their human counterparts (Caplan, 2018). In prioritising consistency, companies have invested in

Suggested Citation (APA): Suzor, N and Nelson, L. (2024, October). *GPT4 v the Oversight Board: Using large language models for content moderation*. Paper presented at AoIR2024: The 25th Annual Conference of the Association of Internet Researchers. Sheffield, UK: AoIR. Retrieved from http://spir.aoir.org.

systems that perform well on average, but comparatively much worse on harder line-calls and less common categories of decisions and content. It is easier for content classifiers to consistently identify explicit, egregious pieces of prohibited content than to understand subtlety, nuance and context (Dias Oliva et al., 2021; Trott et al., 2022). There is more training data available for common types of rule breach in more common languages, meaning errors are likely to disproportionately impact already-marginalised groups. In the moderation of hate speech and abuse, uneven error rates are compounded by the fact that minority groups are disproportionately targeted by threats and harassment and have their own content flagged for review at a disproportionately high rate (Duguay et al., 2018).

For a long time, the limitations of automated content classifiers in dealing with borderline cases have seemed intractable. With the recent expansion in the capabilities and availability of large language models, however, there is reason to suspect that more nuanced automated assessment of content in context may be possible.

Objectives

We aim to develop a chain of prompts that can reliably distinguish between hate speech, content that may be harmful but is not strictly prohibited, and reclaimed language and counterspeech. As sociolegal scholars, we seek to move beyond binary classification tasks, to the more complex task of generating verifiable interpretations of texts and application of rules expressed in natural language. By breaking challenging moderation problems into chains of smaller tasks, and incorporating automated checks to identify errors, we aim to create a system that is consistent, transparent and reliable. Crucially, we explore how it might be possible to use large language models as an alternative to highly specialised content classifiers and generic ‘toxicity’ and ‘safety’ models. We aim to assess the potential of generative models to identify hateful speech in context, rather than focusing predominantly on explicit language, as current classification-based approaches tend to do.

Method

This analysis focuses on the Oversight Board’s decisions under Meta’s hate speech policies. The Oversight Board selects a very small proportion of appeals from Instagram and Facebook and renders detailed decisions that reflect extensive deliberation and external consultation. They arguably represent some of the hardest cases in contemporary content moderation, and they are certainly the most detailed and extensive analyses of their kind with publicly available reasoning.

We deliberately focus on hard cases as a way of exploring the potential that foundation models present to approach content moderation tasks in a radically different way. We selected these cases to explore opportunities to use large language models in ways that more closely resemble and support the analytical tasks of interpreting content and applying the complicated (and sometimes poorly written) natural language rules, definitions, and exceptions of content guidelines. These cases, which require understanding of contextual signs and domain expertise, are often thought to require expert human analysis.

In this exploratory stage, we iteratively develop and refine a set of prompts that together form a chain to apply content policies to examples. We take the Oversight Board decisions as ground truth initially, and inductively explore variations in prompting with the goal of producing a generalisable chain that produces consistently good results across the existing hate speech decisions. In addition to systematic tests with varying prompts and context, we integrate and evaluate promising techniques from other research contexts. We include, for example, 'chain-of-thought' prompting to approximate reasoning (Kojima et al., 2022); breaking complex policies into atomistic steps to avoid getting 'lost in the middle' (Liu et al., 2023) of long prompts; and adding key failed responses as 'few-shot' examples (Brown et al., 2020). A fuller description of and reflection on our development process will follow in the full paper.

Preliminary results

Our evaluation in this project is primarily qualitative. Our exploratory methodology does not lend itself to quantitative testing against large-scale held-out datasets. Many existing content moderation datasets are tuned specifically for stress-testing classification models. Meta's Hate Memes dataset, for example, is full of synthetic examples designed specifically to confound classifiers: explicit hateful messages and harmless gibberish that uses the same form or expression as those messages.

Our preliminary results are promising and revealing. We are interested not just in the number of correct *outcomes*, but whether the outcomes are made understandable in a way that provides some degree of confidence that those outcomes might be supported by valid *reasons* that are logical and verifiably correct. Obviously, the outcomes of the language models themselves are still probabilistic; here we explore how well we can use probabilistic generative models to approximate reasoned decision-making. We look forward to finalising this analysis and presenting a detailed examination of our methodological and theoretical contributions in the coming months.

Future work

We hope that this work will lead to useful methodological and conceptual insights that can guide emerging regulation of digital platforms. The requirements of new legislation, including the Digital Services Act in Europe and the Online Harms Act in the UK, are still underspecified. Our analysis is grounded in theories of justice that highlight not just general values of due process but substantive impact on marginalised groups. By conducting this research independently, we aim to expand the range of potential approaches to moderation at scale, beyond those directly envisaged by platforms and regulators.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Caplan, R. (2018). *Content or context moderation?* [Report]. Data & Society Research Institute. https://apo.org.au/node/203666

Dias Oliva, T., Antonialli, D. M., & Gomes, A. (2021). Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture*, *25*(2), 700–732. https://doi.org/10.1007/s12119-020-09790-w

Duguay, S., Burgess, J., & Suzor, N. (2018). Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. *Convergence*, 1354856518781530. https://doi.org/10.1177/1354856518781530

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (1st edition). Yale University Press.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022, May 24). *Large Language Models are Zero-Shot Reasoners*. arXiv.Org. https://arxiv.org/abs/2205.11916v4

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). *Lost in the Middle: How Language Models Use Long Contexts* (arXiv:2307.03172). arXiv. https://doi.org/10.48550/arXiv.2307.03172

Trott, V., Beckett, J., & Paech, V. (2022). Operationalising 'toxicity' in the manosphere: Automation, platform governance and community health. *Convergence*, 1–16. https://doi.org/10.1177/13548565221111075