Aligning and Auditing Large Language Model (LLM) for Harmful Content Detection for Body Dissatisfaction and Eating Disorders (ED): Rule Development and Validation Process

Pranita Shrestha
pranita.shrestha@monash.edu
Monash University
Melbourne, Australia

Jue Xie
jue.xie@monash.edu
Monash University
Melbourne, Australia

Pari Delir Haghighi
pari.delir.haghighi@monash.edu
Monash University
Melbourne, Australia

Michelle Byrne
michelle.byrne@monash.edu
Monash University
Melbourne, Australia

Roisin McNaney
roisin.mcnaney@unimelb.edu.au
University of Melbourne
Melbourne, Australia

Abstract

The ubiquity of social media has amplified concerns about its impact on users' body image and disordered eating behaviours, particularly for individuals at risk of or experiencing eating disorders (ED). Highly visual platforms like Facebook, TikTok, YouTube, and Instagram influence the perceptions and behaviours of millions, posing particular risks for vulnerable audiences. Current tools for moderating harmful content fail to detect nuanced visual, audio, and text-based cues simultaneously. This study leverages Large Language Models (LLM) and multimodal AI to identify potentially harmful content related to body image and eating disorders with contextual understanding. The main goal is to involve stakeholders in the LLM alignment and auditing process, including eating disorders service providers, experts, researchers, and individuals with lived experience.

CCS Concepts

• **Human-centered computing** → **HCI design and evaluation methods**; **User centered design**.

Keywords

social media, body dissatisfaction, eating disorders, large language model

ACM Reference Format:
Pranita Shrestha, Jue Xie, Pari Delir Haghighi, Michelle Byrne, and Roisin McNaney. 2025. Aligning and Auditing Large Language Model (LLM) for Harmful Content Detection for Body Dissatisfaction and Eating Disorders (ED): Rule Development and Validation Process. In *Proceedings of Conference on Human Factors in Computing Systems (CHI '25).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

CHI '25, 978-1-4503-XXXX-X/18/06

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

1 Research Problem and Motivation

The widespread and increasing presence of harmful content on various social media platforms has significantly contributed to body dissatisfaction and eating disorders (eating disorders). This harmful content can include pro-eating disorder (pro-eating disorders) content, promotion of unhealthy beauty standards, online 'challenges' such as A4 waist [8] or thigh gap challenge [10] and so forth. A recent study [7] reported the alarming impact of TikTok (social media platform) algorithm, which recommended individuals with eating disorders 4343% more toxic eating disorder content, 335% more dieting content and 146% more appearance-oriented content than other average users. The constant bombardment of such harmful and triggering content can create a feedback loop for such vulnerable individuals, further amplifying and reinforcing body image concerns, unhealthy behaviours, and fostering disordered eating behaviours such as extreme dieting, vomiting after meals or purging [6].

The current moderation systems employed by social media platforms such as Facebook, TikTok, and Instagram rely heavily on hashtags, keywords and manual user reporting [1–3]. However, due to the context-specific and evolving nature of such content, these systems often struggle to properly identify such harmful content effectively. Efforts by content creators to bypass existing moderation systems further exacerbate this issue [5]. Due to the increasing prevalence and harmful impact of such content, there is an urgent need to develop more effective methods for identifying it and providing necessary intervention for people at-risk of or experiencing eating disorders.

By leveraging the Large Language Model (LLM) with the human-in-the-loop concept, this study seeks to address the limitations of current automated content moderation. It incorporates the insights and perspectives of diverse key stakeholders (eating disorder service providers, researchers and professionals working in the body image and eating disorders space, and lived experience of eating disorders) into performing LLM-alignment and LLM-auditing. LLM-alignment refers to the process of tailoring the output of LLM to align with human values and predefined rules. This approach aims to design Artificial Intelligence (AI) agents leveraging LLM that are contextually aware, effective and backed by evidence which

ensures the need and sensitivity of the targetted population and creates guardrails against which the LLM has to be audited.

2 Methods

This alignment and auditing approach ensures that an LLM effectively identifies harmful social media content related to body image and eating disorders. This approach involves four key stages: 1) Creation of LLM-alignment and LLM-auditing rule development, 2) Expert validation via Delphi Study, 3) LLM alignment, and 4) Validation and auditing process

2.1 Creation of LLM-alignment and LLM-auditing Rule Development

The rule development is a foundational step in alignment and auditing the LLM for harmful content detection about body image and eating disorders. This phase focuses on capturing diverse perspectives to inform the development of nuanced and context-aware factors in identifying the harmful content of this particular space. This will guide what LLM needs to identify and flag while considering the harmfulness factor.

2.1.1 Methods: The study was divided into two parts:
Interviews with Experts by Profession: We conducted interviews with n=12 experts and professionals from the body image and eating disorder space. The experts by profession provided evidence-based perspectives on what constitutes harmful content and effective prevention strategies.

Focus groups with Experts by Lived Experience: The experts by lived experience involved individuals who had a history of eating disorders. They were recruited through the Butterfly Collective, the Butterfly Foundation's (Australia's leading eating disorders service provider) Lived Experience Network. In total, n=18 participants (14 female, 2 male, and 2 gender-diverse individuals) participated in the focus groups. Participants were given the option to join either mixed-gender or homogenous gender groups based on their comfort. This resulted in five focus groups:

- n=3 focus group for female (n=14 participants)
- n=1 focus group for male (n=2 participants)
- n=1 focus group for gender diverse (n=2 participants)

The experts by lived experience provided firsthand insights into harmful content, triggers and social media experiences.

Results: From the analysis of the interviews and focus groups, we developed 67 LLM-alignment and LLM-auditing rules to guide the identification of harmful content related to body image and eating disorders. Examples of these rules include:

- **Rule 1.1:** Content depicting individuals with emphasized visual ribcages, collarbones, or other bones that suggest underweight bodies.
- **Rule 1.3:** Content promoting or encouraging participation in "size challenges" (e.g., "A4 waist challenge," "hip bone comparison").

These rules served as the foundation for the Delphi study, where experts by profession and lived experience refined and validated its effectiveness for alignment and auditing LLM on a large scale.

2.2 Expert validation via Delphi Study

We are using a Delphi study to achieve expert consensus on identifying harmful social media content related to body image and eating disorders. A Delphi study is a structured and systematic process that is used to gather consensus on expert opinions or certain topics [11]. It is a highly robust methodology that is generally used in healthcare research. Our Delphi study will leverage the knowledge, perspectives and experience of experts by profession (researchers and professionals working in the body image and eating disorders space) and experts by lived experience (individuals with a history of eating disorders). Since the detection and mitigation of harmful content related to body image and eating disorders is complex, nuanced and emerging, the use of the Delphi study can ensure that the LLMs are informed by real-world perspectives and aligns with societal needs and values.

To guide the process, we formed a Steering Committee (SC) comprising n=6 members, including two of the authors. The remaining n=4 members were either experts by profession or experts by lived experience. These four members were the only ones to participate in two rounds of the Steering Committee survey.

The SC reviewed each of the 67 proposed LLM-alignment and LLM-auditing rules individually through a structured survey. For a rule to be included in the Delphi study, it must receive more than or equal to 75% consensus among the SC members. In addition to approving or rejecting rules, the committee could suggest new rules or recommend improvements to the existing ones. Any rules that did not pass in the first round were revised based on feedback and presented in the second round, allowing the committee to reassess and refine their decisions.

Simultaneously, we are planning to recruit n=24 experts by profession and n=24 experts by lived experience. These participants will engage in two rounds of Delphi panel surveys, where they will review and either agree or disagree with the 67 rules. The process for the Delphi panel will mirror the SC survey, including opportunities for panellists to suggest modifications or propose new rules. After each round of the Delphi panel survey, SC members will have a meeting and revise the failed rules based on the feedback received from the panellists.

After completing the second round of the Delphi panel survey, the final set of consensus-driven rules will be established. These rules will serve as the foundation for the design and development of the AI agents incorporating LLM in this research.

2.2.1 LLM alignment. Recent advancements in the LLM field have introduced models with multimodal capabilities, enabling them to interpret images and videos alongside text [4]. This research aims to use LLM-alignment techniques to refine existing LLMs for categorising harmful social media content related to body image and eating disorders.

In simpler terms, we will leverage existing LLMs trained on large datasets to first evaluate their ability to understand social media content and detect patterns associated with body dissatisfaction and eating disorders. Then, we will use LLM-alignment and LLM-auditing rules, and we will guide these LLMs to not merely perform mimicry but to provide outputs that reflect the nuanced identification and categorisation of harmful social media content.

This alignment process will involve prompt engineering, an iterative method of crafting input prompts to guide the LLM's output. While prompt engineering may appear to only guide surface-level behaviour of LLM, however, it can provide more nuanced and foundational support when it is informed and guided by system-level understanding. By tailoring prompts to specific needs and contexts, we can improve the accuracy and relevance of the responses [9]. Importantly, prompt engineering functions not only as a mechanism for behavioural control but also as a diagnostic tool that surfaces the model's implicit biases, strengths, and limitations. The Monash-authorised OpenAI ChatGPT Enterprise version will serve as the LLM for this study, ensuring data security and research-specific use. This way we will audit the existing LLM with no alignment and LLM with alignment.

The goal of this research is to enhance the LLM's understanding of harmful social media content by providing it with additional context about why certain material may be considered damaging. This will improve its ability to categorise social media content related to body image and eating disorders. Through collaboration with a key partner, we have secured a TikTok dataset containing approximately 1 million entries. The dataset includes metadata of TikTok videos (publicly available), participants' TikTok usage patterns, and detailed information about their body image, body dysmorphia, and eating disorders, as measured by the Body Image Scale, Body Dysmorphic Disorder (BDD) scale, and the Eating Pathology Symptoms Inventory (EPSI) scale and subscale. Approval for the use of this dataset has been obtained from the collaborating university, which has also conducted an initial categorisation of the content. This categorisation will serve as the benchmark to evaluate the effectiveness of our LLM alignment.

2.3 Validation and auditing process

After completing the LLM alignment using our developed AI rules, we will test and evaluate its performance by inputting content from our dataset into the model. The LLM will generate text outcomes that explain the context of the content and identify whether it is harmful in terms of body image and eating disorders. These classifications will then be compared to those conducted by the collaborating partner, which serves as the gold standard dataset.

The next step will involve human evaluators from diverse backgrounds (with appropriate precautions) to review and validate the LLM's text outcomes for the social media content. These evaluators will audit whether the classifications generated by the aligned LLM (with rules) align with the gold standard. By incorporating our developed rules, the LLM is expected to classify and contextualize social media content more effectively and accurately.

To audit the LLM's performance (with and without alignment), we will use both quantitative and qualitative methods. Quantitative analysis will measure the alignment between the LLM's outcomes and the expert classifications, while qualitative analysis will explore the quality and contextual accuracy of the generated responses. Additionally, incorporating evaluators from diverse backgrounds will enhance the reliability of the validation process and help minimize bias, while also contributing to the diversity of perspectives within the dataset.

3 Contributions to the field of HCI

This research makes a contribution to the field of Human-Computer Interaction (HCI) by emphasising the importance of involving stakeholders in the guiding and auditing of LLM. It highlights the need to adopt participatory design approaches and methodologies from other fields (Delphi study), where diverse stakeholders including mental health professionals, service providers, researchers and lived experience, are actively engaged throughout the process. By incorporating their perspectives, the research ensures that LLMs are ethically sound, contextually aware, and aligned with the needs of vulnerable populations, such as individuals at risk of or experiencing ED. This collaborative approach helps bridge the gap between technical innovation and human-centred design, fostering systems that are not only effective but also socially responsible.

In addition, the research enhances the HCI's understanding of multimodal interaction design by combining visual, textual and audio content analysis. This aids in providing nuanced and contextual understanding. By using LLM to process LLM-alignment and LLM-auditing input (rules that have received consensus from experts by profession and experts by lived experience), the research also contributes to demonstrating how diverse perspectives can be synthesised to inform LLM alignment and auditing.

References

[1] 2021. How We're Supporting People Affected by Eating Disorders and Negative Body Image. https://doi.org/news/2021/02/supporting-people-affected-by-eating-disorders-and-negative-body-image/

[2] 2024. Help Center. https://doi.org/252214974954612

[3] 2024. Safety Center. https://doi.org/safety/en/eating-disorder

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[5] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society* 20, 12 (2018), 4492–4511.

[6] Sasha Gorrell and Stuart B. Murray. 2019. Eating Disorders in Males. *Child Adolesc Psychiatr Clin N Am* 28, 4 (2019), 641–651. https://doi.org/10.1016/j.chc.2019.05.012

[7] Scott Griffiths, Emily A Harris, Grace Whitehead, Felicity Angelopoulos, Ben Stone, Wesley Grey, and Simon Dennis. 2024. Does TikTok contribute to eating disorders? A comparison of the TikTok algorithms belonging to individuals with eating disorders versus healthy controls. *Body Image* 51 (2024), 101807.

[8] Todd Jackson, Xiaoxuan Ye, Brian J Hall, and Hong Chen. 2021. "Have you taken the A4 challenge?" Correlates and impact of a thin ideal expression from Chinese social media. *Frontiers in Psychology* 12 (2021), 669014.

[9] Eun-young Lee, Ngagaba Gogo Dae Il, Gi-hong An, Sungchul Lee, and Kiho Lim. 2023. ChatGPT-Based Debate Game Application Utilizing Prompt Engineering. In *Proceedings of the 2023 International Conference on Research in Adaptive and Convergent Systems*. 1–6.

[10] Kerrie Caitlin Leonard. 2020. *The Impact of Social Media Body Challenges on Youths' Body Image.* Master's thesis. North Dakota State University.

[11] Prashant Nasa, Ravi Jain, and Deven Juneja. 2021. Delphi methodology in healthcare research: how to decide its appropriateness. *World journal of methodology* 11, 4 (2021), 116.