Building an LLM-Powered Content Moderation Bot in the Classroom

Shelby Grossman, *Arizona State University, USA*
Anthony Mensah, *Stanford University, USA*
Alex Stamos, *Stanford University, USA*
Jeffrey Hancock, *Stanford University, USA*

ABSTRACT As online platforms seek to improve content-moderation strategies, large language models (LLMs) may be a potential tool. This study examines opportunities and limitations of LLM-powered moderation through a unique lens: student projects for a Stanford University course titled Trust and Safety. In this course, students developed Discord bots using LLMs to moderate specific types of harmful content. Interviews with 16 of the students suggest that these models demonstrate high accuracy, often exceeding students' expectations. Notably, in cases of disagreement between the student and the model, closer analysis frequently validated the model's judgments. However, students also observed limitations: LLMs proved unhelpfully sensitive to prompt phrasing and exhibited many contextual interpretation challenges common to human moderators and traditional machine-learning classifiers.

Trust and safety refers to the field of detecting and mitigating harmful uses of the Internet, using products the way that they are designed to work. For example, if a death threat is sent via TikTok's direct-messaging feature, the sender is using the feature as intended. An account has not been hacked, but the content of the message itself is harmful. At technology companies, trust and safety roles might include policy positions to design guidelines to prevent user harassment or scams and engineering roles to develop technologies that detect child sexual-abuse images on a platform.

We teach an interdisciplinary course at Stanford University titled Trust and Safety. Most weeks focus on a specific type of harm, with the aim of helping students to develop empathy for victims, understand how the harm manifests, and examine existing policy and technical responses. Throughout the course, students work in small groups to design and build a content-moderation bot capable of detecting and responding to an abuse type of their choosing. Many groups choose to incorporate large language models (LLMs) into their bots. LLMs are widely predicted to play an important role in content moderation in the future.

The course is designed to encourage students to think about the potential harms that an online platform could enable, particularly for those who someday might create their own online service. The course also aims to deepen understanding of the challenges and tradeoffs associated with implementing safety mitigations.

The experience students have of researching an abuse type and then designing, building, and evaluating a moderation bot in only a 12-week quarter mirrors the process of a small start-up developing a content-moderation system. Start-ups move quickly, often under resource constraints, and the student experience provides a window into the same types of obstacles and opportunities that these companies might encounter: how to (1) make quick decisions balancing enforcement tradeoffs; (2) integrate human oversight with large amounts of data and limited personnel; and (3) iterate rapidly.

Corresponding author: Shelby Grossman *is a professor of practice at the Howard Center for Investigative Journalism at Arizona State University's Walter Cronkite School of Journalism and Mass Communication. She can be reached at sgrossm7@asu.edu.*
Anthony Mensah *is an alum of Stanford University and currently works at Meta. He can be reached at admensah@alumni.stanford.edu.*
Alex Stamos *is a lecturer at Stanford University and chief security officer of Corridor Security. He can be reached at stamos@stanford.edu.*
Jeffrey T. Hancock *is the Harry & Norman Chandler Professor of Communication at Stanford University. He can be reached at hancockj@stanford.edu.*

doi:10.1017/S1049096526101929

To explore their perceptions of LLMs as tools for content moderation, we interviewed 16 students from the class about their experiences in developing bots using LLMs. We report their findings as substantive findings about the opportunities and challenges of using LLMs for content moderation; indeed, we demonstrate that many student observations are consistent with findings in the academic literature. Additionally, the interview responses can be viewed from a pedagogical perspective, showing how students thought through the project and their key takeaways from the experience.

The course is designed to encourage students to think about the potential harms an online platform could enable, particularly for those who someday might create their own online service.

THE TRUST AND SAFETY COURSE

We have been developing and refining a trust and safety course at Stanford University for five years. Initially, the course was a computer science–specific offering titled Trust and Safety Engineering. In the third year, we introduced a companion course in the political science department. In 2024, we taught an interdisciplinary course open to students majoring in computer science, communication, and international policy programs. The 2024 iteration enrolled 161 students: 140 from computer science, 13 from communication, and eight from international policy.

Co-taught by a practitioner with a background in computer science and scholars in communication and political science, the course examines a range of online harms, including misinformation, hate speech, terrorism, harassment, and sexual exploitation. We also included modules on harms that are specific to platforms that enable offline interactions (e.g., ride-sharing) and to content moderator wellbeing.

Co-taught by a practitioner with a background in computer science and scholars in communication and political science, the course examines a range of online harms, including misinformation, hate speech, terrorism, harassment, and sexual exploitation.

The course's primary assessments consist of three milestones. For the first milestone, students work individually to select and research a specific abuse type (e.g., incitement to violence). They then synthesize their findings in a memo to an imaginary CEO describing the dynamics of the harm, including what is known about perpetrators and victims and technological and policy interventions to mitigate this harm.

For the second and third milestones, students collaborate in interdisciplinary groups of five that deliberately mix students who have technical and nontechnical backgrounds. Each group selects a specific harm type and develops a content-moderation bot in Discord. We provided two channels for each team: one that simulates their preferred online platform (e.g., a public feed analogous to X) and another for moderation. Students were required to implement an automated abuse-detection approach that could involve training a classifier, leveraging existing classifiers, using an LLM, or a combination of these methods. We provided starter code to set up the Discord bot, and students took it from there.

Students also must evaluate their bot with a confusion matrix, which is an approach for assessing the extent to which their bot over-enforces (i.e., assesses that content is harmful when it is not, also known as a false positive) or under-enforces (i.e., assesses that content is not harmful when it is, also known as a false negative). For example, one group's bot, designed to moderate doxxing (i.e., revealing someone's address as a harassment tool), failed to recognize a foreign address as an address, which is a false negative. Another group's bot, built to moderate hate speech, incorrectly flagged the sentence "Asian people are the shit" as hate speech, which is a false positive.

Students were asked to calculate precision and recall statistics. Precision is the portion of content flagged as harmful that actually is harmful. Recall is the portion of harmful content that is flagged as harmful. There often is a tradeoff between optimizing for one over the other, and students could decide which to optimize based on, for example, the type of harm they are investigating.

The first milestone, as well as part of the second, can be integrated into courses with only political science students. The first milestone focuses solely on research and the second asks students to design a user-reporting flow and a moderator flow, both of which can be completed without any coding.

We provided each group of students with $50 in Google Cloud Platform credits to deploy their bot. In addition, many students pooled funds to buy OpenAI credits; groups reported that $10 generally was sufficient for the project. Many groups also used the Google Perspective application programming interface (API), which is a free API that provides toxicity scores for text.

At the end of the quarter, students demonstrated their bots at a poster session. We invited trust and safety professionals from industry and Stanford University faculty to serve as guest judges. The poster session also served as an informal networking opportunity; special stickers indicated whether a student was seeking employment or a guest judge was hiring.

To make the project more concrete, we provide an example of an actual student group's project. This group chose to focus on the "pig-butchering" scam. In this scam, perpetrators develop a relationship with victims over weeks or months and then request increasingly larger "investments." The group developed policy language that prohibited this type of scam and outlined enforcement measures. The group then created a dataset, including text

Table 1

GPT-4o Content-Moderator Performance Metrics

	Predicted Not Scam	Predicted Scam
Not Scam	88	6
Is Scam	42	57

Notes: Accuracy: 75%, Precision: 90%, Recall: 58%. To better test their bots, students often worked with datasets that contained artificially high proportions of harmful content. Source: Stanford University student project for Trust and Safety course in 2024.

from screenshots of pig-butchering scam posts (sourced from a subreddit about scams) and innocuous investment-opportunity text (from an investment subreddit). They used the OpenAI GPT-4o API and instructed the model to act as a content moderator with expertise in the pig-butchering scam. Additionally, the group developed a program to check text for suspicious links and implemented a rule to flag users who messaged three or more unconnected individuals within a week. To assess the LLM's performance, the group compared its assessments to the "ground truth" of whether the text came from the scam or the investment subreddit, as shown in the confusion matrix in table 1. Finally, the group outlined the next steps that they would implement with more time to address identified weaknesses in their bot's performance.

Political science studies have theorized about the demand for and consequences of automated moderation of social media content. Gorwa, Binns, and Katzenbach (2020) expressed concerns about transparency and auditability of automated moderation. Pradel et al. (2024) argued that in some contexts, users are skeptical of even the need for content moderation. Considering the anticipated if not current use of LLMs for social media moderation, providing students with the opportunity to design, build, and assess a moderation tool creates informed critics of platform moderation. Ideally, the project makes the students (who may go on to all types of technology and government policy jobs) experts in the potential benefits and limitations of AI-driven moderation.

METHODS

For our research, the goal was to interview as many students as possible who used AI in their content-moderation bots, aiming to understand their thought process and lessons learned. To study Stanford students, we followed a specific university review process. First, we submitted our proposed research to the university's Student Data Oversight Committee. This committee recommended research protocols and then signed off on our plan before it was processed by the Stanford University Institutional Review Board (IRB).

Because we began participant recruitment when the course was already ongoing, a colleague who was not involved in the class emailed students to invite them to participate. This ensured that students would not feel pressure or expect that their participation would impact their grade. They were informed that their willingness to participate would be unknown to the teaching team until after final grades were submitted. At that time, a member of the teaching team followed up with an additional recruitment email.

Of the 161 students in the class, 16 agreed to be interviewed: 12 from computer science, one from communication, and three from international policy programs. The interviewed group consisted of 10 undergraduate and six graduate (primarily master's) students, distributed across 12 student groups. Our sample included a higher proportion of women and non–computer science students compared to the full class: women constituted 44% of our interviewed students but comprised only 37% of the entire class. Non–computer science students represented 25% of our sample but only 13% of the full class.

The groups investigated the following harms:

- pig-butchering scams
- animal-abuse photos
- terrorism promotion
- violence incitement
- cryptocurrency scams
- catfishing
- adversarial coordinated reporting
- false information
- cyberbullying
- hate speech

Of the 12 groups, eight used an OpenAI model, three used Gemini, one used Claude, one used the Cohere Command R+ model, one used LLaMa v1, and one used Mistral. (Some groups tested multiple models.)

We conducted semi-structured 30-minute interviews. Fifteen of the interviews took place in July 2024, one month after the quarter had ended, with an additional interview in October 2024.

FINDINGS

This section describes findings from the student interviews.

Prompting

Students took different approaches in their requests of the LLMs. Many groups tasked the LLM with making binary decisions (i.e., yes or no), such as determining whether a post violated a content policy. Other groups asked the LLM to classify the harm type, compare two posts to assess which was more harmful, or rank a post's harmfulness on a scale (e.g., from 1 to 5). Some groups also requested that the LLM provide a confidence assessment of its decisions. One group used an LLM to generate customized mental health resources for users based on the content of their posts.

Students experimented with different types of prompts to increase bot accuracy. They found that models often included extraneous information in their assessment of content,[1] and their outputs were sensitive to subtle changes in prompt wording. Additionally, students faced a tradeoff in using short or long prompts. Although longer prompts often improved detection for specific harms, they frequently reduced the model's reasoning capabilities. One student observed that when instructing the LLM on a content policy, one must "either specify everything or nothing."[2] For example, if the policy includes an example of bullying, every possible form of bullying must be listed or risk that the model enforces the policy only for the explicit examples provided.

Many groups had success with using chain-of-thought logic in the prompt,[3] in which the students "tell the model to reason in order, step by step" (Willner and Chakrabarti 2024).[4] For example, one group wanted to moderate harmful falsehoods. It asked the

model to first assess whether a statement was false. If it were false, the group then requested the model to assess whether the content could be harmful.[5]

Groups that tested multiple LLMs discovered that a prompt effective for one model may not work for another. Some models, for example, would not permit prompts to contain certain harm-related keywords[6] and others refused to assess certain types of harmful content.[7]

Harms That May Not Be in the Training Set

We interviewed two students from different groups that focused on building a bot to detect pig-butchering scams. One student sourced pig-butchering scam messages from targets who had shared their experiences on a scam-focused subreddit. The other student's group used GPT-4o to generate synthetic chat examples, which they then asked the model to analyze and assess for potential scam characteristics.

One student observed that GPT-4o could accurately define a pig-butchering scam but struggled to apply that definition to specific examples.[8] The student noted that much of the conversation in a pig-butchering scam resembles a casual exchange between friends, adding that the model often "overlooked the full context of the conversation."

One student noted that toward the end of the scam, when the scammers request an "investment" or financial assistance, they often encourage the target to transition to Telegram or another platform with minimal moderation.[9] This student speculated that because LLMs are unlikely to have been trained on Telegram direct messages, they likely would not recognize the specific scam text commonly used on this platform. Moreover, if key parts of the conversation occurred off of the platform that is attempting the moderation, detection would become even more challenging. One student explained that the model had been prompted with the anticipated timeframe for when the "ask" would occur; however, incorrect predictions on timing, of course, would decrease model accuracy.[10]

Another challenge involved over-enforcement by the LLM. One group recognized that a key element of the scam was an appeal to kindness, and it incorporated this insight into its prompt to the LLM. However, this approach led to many false positives in which the model flagged innocuous chats.[11]

Despite these challenges, one student pointed out that pig-butchering scammers often reuse identical text.[12] If LLMs could be trained on this text, detecting these scams might not be particularly difficult, even though the conversations often appear indistinguishable from regular chats.

A final issue related to the financial cost of evaluating these lengthy and often months-long conversations with an LLM. To address this, one group used the bidirectional encoder representations from transformers (BERT) model (a non-generative LLM designed for text analysis) to make an initial assessment of whether a conversation might be a scam. Only those conversations that were flagged as potentially suspicious were sent to the LLM for further evaluation. However, this approach had limitations because the BERT model often struggled to make accurate assessments.[13] The academic literature similarly has proposed clever strategies for reducing the amount of content that a model reviews. Qiao et al. (2024), for example, proposed clustering content and then asking a model to review a representative piece of content from a cluster.

Knowledge of Current Events

Students' interviews highlighted that their LLM bots faced challenges reflective of broader, well-documented content-moderation issues. One student whose group focused on incitement to violence noted that GPT-4o struggled to moderate content requiring an understanding of recent events, which the model lacked. The student cited an example of a statement by US President Joe Biden one week before an assassination attempt on President Donald Trump. Biden had encouraged Democrats to put Trump in a "bullseye" (Nelson 2024), a statement that would have a different meaning after the assassination attempt.[14] In Spring 2024, LLMs generally did not have knowledge of events that occurred after their training cutoff and could make decisions that did not reflect current events, just as humans might.

Perceptions of LLM Success

We asked students to assess how successful they believed their LLM bot had been. The consensus was that LLMs outperformed non-LLM approaches, particularly excelling in handling difficult edge cases.

Even in cases in which the LLM produced what initially appeared to be a false positive or false negative, students observed that closer examination often revealed that the LLM might have been "right" whereas the supposed ground truth was "wrong." As one student explained, "A lot of times we looked at the model and realized that even though it was a different response than what we labeled, the model was also right."[15] This observation aligns with examples in the academic literature, in which scholars similarly concluded that the model's output proved correct on further review (Kumar, AbuHashem, and Durumeric 2024).

Extant literature frequently examines whether models are better at precision or recall (Kumar, AbuHashem, and Durumeric 2024; Zhan et al. 2024), and we asked students to conduct similar analyses for their projects. Among the projects for which we had permission to analyze results and that clearly reported precision and recall statistics, two thirds showed higher precision and one third showed higher recall. Academic research similarly found that LLMs used for content moderation are better at precision than recall (Kumar, AbuHashem, and Durumeric 2024). However, it is unlikely that students interpreted their findings to mean that LLMs are inherently better at one metric compared to the other. Instead, they tailored their moderation prompts to prioritize either precision or recall based on the specific harm type that they wanted to address. For example, a group that focused on cyberbullying explicitly prioritized precision over recall, thereby not overworking hypothetical human moderators.

A potential pitfall in the students' process was overfitting, in which (for example) an LLM is trained on policy language so specific that its performance excels on a particular dataset but may decline when applied to a wider range of data. One group explicitly addressed this in its project, noting that it deliberately constrained the model fine-tuning process to avoid overfitting. However, the students' reported precision and recall rates, as well as their overall impression of LLM performance, might have been influenced by overfitting. Actual performance on a broader dataset, therefore, could be less impressive.

Potential for Adversarial Behavior

We asked students for their thoughts on how an adversarial actor might respond if they knew an online platform was using an LLM for content moderation. Several students suggested that a bad

actor with access to the same LLM could test its limitations off-platform, although this strategy may not work without knowledge of the prompt or access to a particular fine-tuned model.[16] Another student proposed that the adversarial use of video-game language could evade some LLM safeguards.[17]

Are LLMs the Future of Content Moderation?

We asked students whether their experience with the project left them feeling optimistic or pessimistic about the potential use of LLMs for content moderation. Six students expressed optimism, one was pessimistic, and nine had mixed feelings.

Among the optimistic students, several expressed hope that LLMs would reduce the need for human moderators to review distressing material.[18] These students also were impressed with how well the LLM performed, given that they spent only a few weeks on the LLM portion of the project.[19] They noted that the model worked well despite relying on fairly short prompts.[20] With more time to experiment with prompts, many believed its accuracy could be improved even further.

The optimistic students also emphasized the potential for LLMs to complement existing moderation systems. They believed that LLMs would be most effective when integrated with processes such as user reports,[21] user appeals,[22] and human review.[23]

Students raised several LLM limitations. One student stated, "There has to be a person at least at [some] stage. I wouldn't trust a model with all of this."[24] Another student said that they would "be wary of making decisions entirely based on LLM unless the confidence level is very high."[25]

One student noted that although LLMs struggle with cultural nuance, a point also made by Narayanan and Kapoor (2024), human moderators often face similar challenges.[26] Another student observed that LLMs might excel in moderating straightforward harmful content (e.g., single messages about scams or explicit images) but could be less effective for more complex cases requiring contextual understanding.[27] To improve performance, students recommended combining LLM evaluation of a post with user-profile data, post comments, and other contextual information,[28] the latter being a point generally supported in academic literature (Kumar, AbuHashem, and Durumeric 2024; Moon et al. 2023).

Two students voiced their concerns about the resource intensity of using LLMs for content moderation, citing the slow response times and financial costs,[29] opinions also expressed by Willner and Chakrabarti (2024). One student expressed skepticism about the ability of LLMs to outperform machine-learning models developed by large social media companies. However, this student suggested that LLMs could benefit smaller companies that lacked those resources.[30]

Another issue centered on the static nature of AI models. One student was concerned that relying heavily on LLMs could entrench outdated understandings of harmful content,[31] which also was discussed by Narayanan and Kapoor (2024).

Some students reflected on broader challenges. One observed that platforms uninterested in moderation would not benefit from LLM advancements. For those platforms, "no LLM technology under the sun will help. You can lead a horse to water, but at the end it's up to them whether to take advantage of it."[32] This student's group had tested their LLM on posts from the under-moderated platform 4chan.

A student currently working to incorporate an LLM into a company's trust-and-safety pipeline shared insights from this professional experience. The student noted that because the work faces resistance from trust-and-safety colleagues, strategies to make the technology more palatable were used, framing the LLM's role as "enhancing" instead of "replacing" other moderation tools.[33]

The one student who was pessimistic noted "a lot more upside for the bad actor [...] than for good actors."[34]

CONCLUSION

Many students were impressed with the performance of their LLM-powered content-moderation bot. However, the project reinforced a lesson that is not new but remains important: automated content moderation inevitably results in both false positives and false negatives, and automated tools can be most effective when combined with other forms of moderation, including human review. The optimal balance between automation, human review, and other moderation tools should depend on the specific type of abuse being moderated and the potential harms associated with over- and under-enforcement.

...the project reinforced a lesson that is not new but remains important: automated content moderation inevitably results in both false positives and false negatives, and automated tools can be most effective when combined with other forms of moderation, including human review.

ACKNOWLEDGMENTS

This study was approved by Stanford University's IRB Protocol #75456. We thank the students who participated in this research, as well as the other students in the class and the course assistants.

CONFLICTS OF INTEREST

The authors declare that there are no ethical issues or conflicts of interest in this research. Shelby Grossman occasionally consults for OpenAI as part of its external red teaming network. ■

NOTES

1. Interview with Student J on July 22, 2024.
2. Interview with Student G on July 18, 2024.
3. Interviews with Student O on July 29, 2024; and Student P on July 29, 2024.
4. Students read Willner and Chakrabarti (2024) during class. It is possible that learning about the efficacy of chain-of-thought prompting influenced their perception of its usefulness for their project.
5. Interview with Student P on July 29, 2024.
6. Interviews with Student D on July 17, 2024; and Student L on July 24, 2024.

7. Interview with Student O on July 29, 2024.
8. Interview with Student A on July 16, 2024.
9. Interview with Student A on July 16, 2024.
10. Interview with Student M on July 25, 2024.
11. Interview with Student M on July 25, 2024.
12. Interview with Student A on July 16, 2024.
13. Interview with Student M on July 25, 2024.
14. Interview with Student D on July 17, 2024.
15. Interviews with Student O on July 29, 2024; and Student P on July 29, 2024.
16. Interviews with Student H on July 18, 2024; and Student J on July 22, 2024.
17. Interview with Student L on July 24, 2024.
18. Interviews with Student A on July 16, 2024; Student H on July 18, 2024; and Student J on July 22, 2024.
19. Interview with Student A on July 16, 2024.
20. Interview with Student B on July 16, 2024.
21. Interview with Student G on July 18, 2024.
22. Interview with Student F on July 17, 2024.
23. Interview with Student Q on October 31, 2024.
24. Interview with Student G on July 18, 2024.
25. Interview with Student J on July 22, 2024.
26. Interview with Student N on July 25, 2024.
27. Interview with Student E on July 17, 2024.
28. Interview with Student O on July 29, 2024.
29. Interviews with Student P on July 29, 2024; and Student L on July 24, 2024.
30. Interview with Student O on July 29, 2024.
31. Interview with Student D on July 17, 2024.
32. Interview with Student C on July 16, 2024.
33. Interview with Student I on July 19, 2024.
34. Interview with Student M on July 25, 2024.

REFERENCES

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7 (1): https://doi.org/10.1177/2053951719897945.

Kumar, Deepak, Yousef AbuHashem, and Zakir Durumeric. 2024. "Watch Your Language: Investigating Content Moderation with Large Language Models." In *Proceedings of the International Association for the Advancement of Artificial Intelligence Conference on Web and Social Media* 18:865–78.

Moon, Jihyung, Dong-Ho Lee, Hyundong Cho, Woojeong Jin, Chan Young Park, Minwoo Kim, Jonathan May, Jay Pujara, and Sungjoon Park. 2023. "Analyzing Norm Violations in Live-Stream Chat." https://arxiv.org/abs/2305.10731.

Narayanan, Arvind, and Sayash Kapoor. 2024. *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*. Princeton, NJ: Princeton University Press.

Nelson, Steven. 2024. "Biden Defends 'Bullseye' Remark About Trump During NBC Interview with Lester Holt: 'I Didn't Say Crosshairs.'" *New York Post*, July 15.

Pradel, Franziska, Jan Zilinsky, Spyros Kosmidis, and Yannis Theocharis. 2024. "Toxic Speech and Limited Demand for Content Moderation on Social Media." *American Political Science Review* 118 (4): 1895–912.

Qiao, Wei, Tushar Dogra, Otilia Stretcu, Yu-Han Lyu, Tiantian Fang, Dongjin Kwon, Chun-Ta Lu, Enming Luo, Yuan Wang, Chih-Chun Chia, Ariel Fuxman, Fangzhou Wang, Ranjay Krishna, and Mehmet Tek. 2024. "Scaling up LLM Reviews for Google Ads Content Moderation." In *Proceedings of the 17th Association for Computing Machinery International Conference on Web Search and Data Mining*, WSDM '24, 1174–75.

Willner, Dave, and Samidh Chakrabarti. 2024. *Using LLMs for Policy-Driven Content Classification*. Austin, TX: Tech Policy Press. www.techpolicy.press/using-llms-for-policy-driven-content-classification.

Zhan, Xianyang, Agam Goyal, Yilun Chen, Eshwar Chandrasekharan, and Koustuv Saha. 2024. "SLM-Mod: Small Language Models Surpass LLMs at Content Moderation." https://arxiv.org/abs/2410.13155.