

AI Content Moderation in Therapy Conversations

Jiwon Kim, Claire Wang, Taeung Yoon, Sabelle Huang, Koustuv Saha

University of Illinois Urbana–Champaign
United States

{jiwonk7, claire46, tuyoon2, sabelle2, ksaha2}@illinois.edu

Abstract

Large language models (LLMs) are increasingly being used for emotional support. They are also being developed for formal therapy purposes. However, LLMs like ChaptGPT or Llama are often developed with content moderation guardrails that prevent them from discussing sensitive subjects with users for both liability and safety purposes, and this inability to broach these subjects may affect their capacity as therapists. In this study, we perform an algorithm audit on three state-of-the-art moderation systems (OpenAI’s moderation endpoint, Meta’s Llama Guard, and Google’s Shield Gemma) to investigate the extent to which these systems flag the content of real-life therapy sessions as undesirable. Our results raise implications for the limitations that users and organizations may encounter when designing LLMs to play the part of therapist.

Introduction

Mental health remains a pressing public health challenge. In the U.S. alone, more than one in five adults live with a mental illness, and many face persistent barriers to accessing timely and affordable care (NIH 2024). At the same time, loneliness and emotional isolation have been linked to a range of serious health risks, underscoring the broader importance of accessible forms of support (eClinicalMedicine 2023). As a result, people increasingly turn to online and digital systems—including online forums, chatbots, and large language model (LLM)-based agents—for emotional support, advice, and companionship (Sharma et al. 2024; De Choudhury, Pendse, and Kumar 2023).

To assuage their feelings of isolation, users are employing platforms such as Replika and ChatGPT to serve as emotional supports and even therapists (Li et al. 2025a; Yuan et al. 2026). There has also been a growing area of research around designing LLM-based conversational agents to fulfill these roles (Fang et al. 2025; Nie et al. 2025; Kim et al. 2026). Many of these agents, such as ChatGPT, Gemini, and Llama, are designed with content moderation guardrails (Markov et al. 2023) that prevent the conversational agent from broaching subjects that may be considered sensitive or explicit, to prevent the agent from causing harm to the user. Content moderation has become a

fundamental aspect of video-sharing platforms, social media sites, and online forums, to facilitate productive discussions and prevent unchecked vitriol toward other users. In the case of conversational agents, many more sensitive topics that are discussed in human-to-human therapy sessions, such as suicidal ideation or thoughts of self-harm, may be flagged by the content moderation system as inappropriate content, thereby preventing the conversation from continuing. In real-life therapy sessions, these topics are integral to discuss so that they can be parsed and processed in a healthy, controlled environment. However, this presents a possible constraint of these conversational agents in their ability to provide mental health support: if they cannot discuss these topics with users, how well can they take on roles to support their mental wellbeing? In this paper, we investigate these limitations.

RQ: How effectively do existing LLM moderation systems distinguish between appropriate and inappropriate content in therapist–client conversations?

To address this question, we adopt Mahomed et. al’s methodology for auditing ChatGPT’s content moderation endpoint (Mahomed et al. 2024). Building off their pipeline, we annotate a dataset of 154 high quality and 104 low quality therapy transcripts (Pérez-Rosas et al. 2019) to provide a ground truth for our audit using OpenAI’s content moderation guidelines. Along with probing OpenAI’s content moderation endpoint to assess the alignment of its flagging of violating content against the human-annotated labels, we also compare the content moderation endpoint’s performance against other state-of-the-art content moderation systems, namely Meta’s Llama Guard 3 (AI 2024) and Google’s Shield Gemma (Google 2024).

We find that all three content moderation systems have low recall, with Omni Moderation at 0.09, Shield Gemma at 0.03, and Llama Guard at 0.34, and similar balanced accuracy scores of 0.53 (Omni Moderation), 0.51 (Shield Gemma), and 0.51 (Llama Guard). These results, especially the low recall across the board, suggest that these content moderation systems underflag the harms in mental health-related conversations, and while they may be able to carry out a conversation on mental health issues, they cannot identify the potential harms in those conversations, and thus may continue them past a safe context with vulnerable users. This lack of sensitivity to such harms suggests that these conver-

sational agents are not yet ready to be deployed in mental health settings.

Finally, we fine-tune a Mistral model using the CounselBench dataset (Li et al. 2025b) to identify potentially harmful language in AI conversational agents. The fine-tuned Mistral model achieved a recall of 0.51 and a balanced accuracy of 0.58, an improvement from the previously assessed content moderation systems. While these scores do suggest that smaller, fine-tuned models may supply an alternative to these general-purpose content moderation systems, we find that even these fine-tuned models miss a large swathe of potentially harmful content, and are thus not yet ready to deploy in the mental health context. This provides additional insights into future attempts to make mental health resources more accessible.

Related Work

Mental health issues such as depression, anxiety, and suicidal thoughts continue to grow worldwide, while access to qualified mental health professionals remains severely limited (Pham, Nabizadeh, and Selek 2022; Wainberg et al. 2017). This imbalance, along with the persistent stigma surrounding mental illness, has encouraged individuals to seek help in digital environments. Online forums, peer-support platforms, and mental health communities have become important alternatives where people can express emotions, share experiences, and receive social validation beyond traditional therapy settings (De Choudhury and De 2014).

With the emergence of large language models (LLMs) and generative AI, conversational agents have begun to occupy a similar role—offering continuous, personalized, and low-barrier emotional support (Chen et al. 2020; Saha et al. 2025; Yoo et al. 2025). These systems have sparked interest in their potential to supplement or even replace aspects of human counseling. Yet, unlike human therapists, LLMs are heavily regulated by built-in content moderation pipelines and “guardrails” that restrict discussion of sensitive or high-risk topics such as self-harm, trauma, and abuse (Mahomed et al. 2024). While these restrictions are crucial for safety and compliance, they can also prevent models from responding to users’ genuine emotional needs, particularly in situations that require acknowledgment of distress or exploration of difficult emotions (Li et al. 2025b).

Recent research has begun to examine how these moderation systems affect user experience and therapeutic quality. Studies suggest that LLMs may produce overly cautious or dismissive replies, interrupting empathic engagement or therapeutic rapport (Kang and Reynolds 2024). Other works raise ethical concerns regarding the opaque decision-making of moderation algorithms and the tension between user autonomy, liability avoidance, and emotional responsiveness (Shi et al. 2026). Recent research has also explored whether LLMs themselves can assist human moderation workflows, pointing to their promise as flexible moderation aids while also raising questions about reliability, domain sensitivity, and deployment context (Kolla et al. 2024; Kumar, AbuHashem, and Durumeric 2024; Zhan et al. 2025). In parallel, recent algorithm audit literature has provided systematic frameworks for identifying algorithmic

T: So Jane it sounds like you’re angry with your supervisor, would I be right in saying that?
C: Well I think I’m pretty angry.
T: All right, maybe to help me understand more about that anger, I’m wondering if you could choose two or three pictures here that look like how you were angry at the time ...
C: Well she got really angry at me because on the way to my room my idiot sister got in the way, so I pushed her out of the way ...

Figure 1: Example of a flagged conversation between a Therapist (T) and a Client (C).

bias and distortion across domains such as search and advertising (Bandy 2021).

Building on this line of inquiry, our study explores how ChatGPT’s content moderation system interprets and flags authentic therapy transcripts, aiming to reveal how safety guardrails shape—and sometimes constrain—the communicative capacity of LLMs designed for mental health support.

Data and Methods

Our study aims to answer whether large language models (LLMs) triggered safety guardrails when processing therapy conversations between human therapists and patients, particularly when sensitive topics were discussed. To address this question, we analyzed the outputs of the OpenAI Moderation model (omni-moderation) (OpenAI 2025) across therapy transcripts, and compare them against human-generated labels. In all, there were 154 high-quality transcripts and 104 low-quality ones. We manually annotate all transcripts to serve as the ground truth and evaluate each transcript for content flags in categories such as harassment, hate, illicit content, self-harm, sexual content, and violence, as defined by OpenAI’s moderation endpoint guidelines (OpenAI 2025). All data was anonymized, ensuring compliance with ethical research standards and protection of individual privacy.

Dataset

The data used in this project were drawn from the counseling conversations collected by Pérez-Rosas et al. (Pérez-Rosas et al. 2019), who investigated the attributes of effective counseling. The dataset is comprised of 258 chat transcripts labeled as high or low quality, assessing the counselors’ behaviors through the framework of the client-centered therapy principles outlined by Miller and Rollnick (Miller and Rollnick 2013). Conversations in which the counselor focused on the client’s perspective and demonstrated empathy were considered high quality, while those in which the counselor primarily provided directives or potentially harmful advice, to which the client complied, were categorized as low quality. In some low-quality conversations, therapist and client labels appeared to be swapped; however, these labeling inconsistencies did not affect the overall content of the conversation, so the original labels were preserved.

Manual Annotation

Our manual annotations of the transcripts served as the ground truth for determining whether a conversation should be flagged for sensitive content. Two authors manually annotated the first 25 low-quality transcripts for the categories of harassment, hate, illicit activity, self-harm, sexual content, and violence, as defined by the OpenAI moderation endpoint (OpenAI 2025). To evaluate the consistency of our manual harm annotations, we computed Cohen’s kappa (κ) between two human annotators. Cohen’s kappa measures inter-rater agreement beyond chance for categorical labels ($\kappa = 1$ indicates perfect agreement, while $\kappa = 0$ indicates chance-level agreement). Across the annotated data, the two annotators achieved $\kappa = 0.64$, suggesting substantial agreement and indicating that the annotation procedure produced reasonably consistent labels between annotators. After more extensive conversation regarding the annotation criteria, the remaining transcripts (both high and low quality) were then divided between the two annotators for individual labeling.

Audit and Finetuning

Using the manual annotation of the transcripts, we calculated its alignment with the outputs of OpenAI’s moderation endpoint, a hosted classifier which assesses the conversation across multiple harm categories and returns both binary flags and per-category severity scores (“category_scores”), using the categories defined above; if any category exceeds the model’s internal threshold, the conversation is marked as “flagged.” To broaden our analysis, we also compare content moderation performance across two other moderation systems: Meta’s Llama Guard 3 (AI 2024), and Google’s Shield Gemma (Google 2024). All prompts and inputs are standardized to ensure each system received identical text; this ensures that the differences in outputs reflected true variations in model behavior rather than input phrasing or formatting. Figure 1 shows a portion of a flagged conversation. Although the APIs do not provide span-level rationales indicating the exact words that triggered a flag, a conversation is flagged when at least one category crosses the model’s decision threshold.

In addition to using existing content moderation APIs, we finetune an open-source Mistral-7B model to detect therapist responses that are overtly harmful, dismissive, aggressive, or otherwise inappropriate for mental health support contexts. Our goal is to examine whether a smaller, domain-adapted language model can more effectively identify toxic counseling behaviors compared to general-purpose moderation systems, following previous research on the effectiveness of SLMs in content moderation (Zhan et al. 2025; Goyal et al. 2025). We train this classifier with the CounselBench dataset (Li et al. 2025b), a large-scale, clinically grounded benchmark for evaluating open-ended mental health question answering. This finetuned model allows us to empirically compare the effectiveness of small, specialized models against widely deployed moderation APIs in mental health settings.

Results

Therapy conversations naturally involve sensitive topics; our goal was to assess whether the LLM detected potentially harmful content appropriately and distinguished between therapeutic discussions and genuinely unsafe material. For this study, a conversation was considered flagged if any part of the transcript triggered any one of the moderation categories described above. We hypothesized that low-quality counseling sessions would yield higher harm values than high-quality ones, as ineffective counseling could fail to mitigate distress and exacerbate sensitive topics.

Importantly, the moderation systems’ outputs are non-deterministic, meaning that repeated runs on the same text can yield slightly different results. This variability arises from the model architecture, infrastructure, and ongoing updates to model weights and snapshots (OpenAI 2024). To mitigate inconsistencies in our analysis, we ran the model on each set of transcripts multiple times, similar to the approach used by (Mahomed et al. 2024).

Moderation Model Results

We analyzed a total of 154 high-quality and 104 low-quality conversations using three moderation models: OpenAI’s omni-moderation-latest (OpenAI 2025), Meta’s Llama Guard 3 (AI 2024), and Google’s ShieldGemma (Google 2024). Each model outputs a transcript-level safety judgment using its own taxonomy.

Across our annotated dataset, OpenAI moderation flagged 2% of annotations as unsafe, and ShieldGemma flagged 1%, yielding broadly similar unsafe rates. In contrast, Llama Guard 3 flagged 28% of annotations as unsafe. This difference is largely explained by Llama Guard 3’s inclusion of a distinct “Specialized Advice” category, which treats specialized financial, medical, or legal advice as unsafe content (AI 2024). In our results, approximately 90% of Llama Guard 3 unsafe annotations fell into this Specialized Advice category. Because the dialogues in the dataset often include therapists offering suggestions and guidance, Llama Guard 3’s taxonomy is more likely to mark clinically routine exchanges as unsafe, inflating the overall flag rate relative to the other two models.

To assess intra-model consistency, we ran each model on each conversation five times and aggregated the outputs using majority vote for the final unsafe/flagged decision. Only a small number of cases varied across runs, typically by adding or removing a single secondary category while preserving the overall safe/unsafe decision.

Finetuned Model Results

Table 1 reports recall and balanced accuracy for detecting toxic therapist responses, comparing multiple moderation systems against human-labeled ground truth. In this evaluation, responses flagged by each moderation model were aligned with annotations indicating whether a therapist’s utterance was toxic or inappropriate in a counseling context. Among the evaluated systems, Omni Moderation achieves a recall of 0.0860 with a balanced accuracy of 0.5260, while

Table 1: Recall and balanced accuracy for detecting toxic therapist responses. The finetuned Mistral model achieves the highest performance on both metrics.

Model	Recall	Balanced Acc.
Omni Moderation	0.0860	0.5260
Shield Gemma	0.0286	0.5143
Llama Guard	0.3429	0.5089
Our Finetuned Mistral	0.5143	0.5821

Shield Gemma performs worse in recall (0.0286) with a similar balanced accuracy of 0.5143. Llama Guard improves recall to 0.3429 but maintains a comparable balanced accuracy of 0.5089. In contrast, our finetuned Mistral judge achieves the highest recall (0.5143) and balanced accuracy (0.5821), outperforming both general-purpose moderation APIs and safety-focused models such as Shield Gemma and Llama Guard.

Despite its smaller size, the finetuned Mistral model achieves the highest recall among all evaluated systems (0.5143), substantially outperforming Llama Guard (0.3429), Omni Moderation (0.0860), and Shield Gemma (0.0286). This improvement in recall indicates that our model is more effective at identifying toxic or clinically inappropriate responses, which is particularly critical in safety-sensitive mental health contexts. At the same time, the absolute recall values remain moderate across all models, suggesting that current moderation systems—whether API-based or safety-focused—fail to detect a significant portion of harmful responses. These results indicate that manually finetuning a smaller model with domain-specific data can be more effective than relying on larger, general-purpose moderation systems, and further highlight the need for additional supervision mechanisms beyond surface-level toxicity detection.

Discussion and Conclusion

There are several approaches that could be additionally explored to overcome or adjust LLM safety guardrails, such as prompt engineering, few-shot learning, or fine-tuning open-source models. Prompt engineering can be used to reframe user inputs or system instructions in ways that reduce refusal behaviors while preserving task relevance. Few-shot learning provides carefully curated examples that guide the model toward desired response styles, potentially steering it away from overly conservative safety responses. Fine-tuning open-source models enables direct modification of model behavior using domain-specific datasets, allowing practitioners to tailor safety boundaries and response characteristics more precisely. Recent research (Song et al. 2025) has shown that machine unlearning can be used to break safety alignment by making an LLM forget its mechanisms for rejecting harmful instructions. These methods are potential directions for future work.

At the same time, directly weakening safety guardrails introduces significant risks, particularly in sensitive domains such as mental health. Rather than removing safeguards, an alternative direction is to complement existing guardrails

with structured, domain-specific supervision mechanisms that enable more nuanced control over model behavior. For example, integrating rubric-based evaluation frameworks or iterative critique-and-revision pipelines may allow systems to remain aligned with high-level safety policies while improving their ability to handle complex, context-dependent interactions (Kim et al. 2026; Goel et al. 2026). Exploring how such approaches can balance flexibility, safety, and clinical appropriateness remains an important direction for future research.

Overall, we found that LLMs alone are not enough to fulfill the role of a therapist. In general, the models did a poor job in determining whether a conversation should be flagged for sensitive content. It is unrealistic to expect the AI to catch every coded or indirect expression, so an LLM model alone could not replace a therapist.

However, in real-world therapy settings, people are usually more straightforward when discussing their emotions, so the risk of such bypassing language affecting AI responses is likely minimal. Still, this prompt sensitivity should be acknowledged as one of the study’s limitations.

We explored whether moderation behavior could be made more comparable across systems by testing models with other models’ taxonomies and guidelines (e.g., evaluating the same transcript with one model while providing the category definitions and decision rules from another). In practice, this did not produce meaningful or reliable results. Classifier-style moderation endpoints (e.g., omnimoderation (OpenAI 2025)) did not adopt external taxonomies beyond their fixed label spaces. As a result, these cross-instruction trials did not improve interpretability or alignment between models’ unsafe judgments, and we treat them as an empirical takeaway about the limits of prompt-based harmonization rather than a basis for our main analyses.

References

- AI, M. 2024. Llama Guard 3: Safety Classifier and Moderation Model. <https://www.llama.com/docs/model-cards-and-prompt-formats/llama-guard-3/>. Accessed: 2025-10-12.
- Bandy, J. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–34.
- Chen, Z.; Lu, Y.; Nieminen, M. P.; and Lucero, A. 2020. Creating a Chatbot for and with Migrants: Chatbot Personality Drives Co-Design Activities. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 219–230.
- De Choudhury, M.; and De, S. 2014. Mental Health Discourse on Reddit: Self-Disclosure, Social Support, and Anonymity. In *Eighth International AAAI Conference on Weblogs and Social Media*.
- De Choudhury, M.; Pendse, S. R.; and Kumar, N. 2023. Benefits and harms of large language models in digital mental health. *arXiv preprint arXiv:2311.14693*.
- eClinicalMedicine. 2023. The epidemic of loneliness. *eClinicalMedicine*, 66: 102395.

- Fang, A.; Chhabria, H.; Maram, A.; and Zhu, H. 2025. Practicing Stress Relief for the Everyday: Designing Social Simulation Using VR, AR, and LLMs. *arXiv:2410.01672*.
- Goel, D.; Lee, J.; Zhong, Q. J.; Rodriguez, V. J.; Brown, D. S.; Karkar, R.; Yoo, D. W.; and Saha, K. 2026. RubRIX: Rubric-Driven Risk Mitigation in Caregiver-AI Interactions. *arXiv preprint arXiv:2601.13235*.
- Google. 2024. ShieldGemma Model Card. https://ai.google.dev/gemma/docs/shieldgemma/model_card_2. Accessed: 2025-10-12.
- Goyal, A.; Zhan, X.; Chen, Y.; Saha, K.; and Chandrasekharan, E. 2025. Momoe: Mixture of moderation experts framework for ai-assisted online governance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 12656–12671.
- Kang, R. M.; and Reynolds, T. L. 2024. "This App Said I Had Severe Depression, and Now I Don't Know What to Do": The Unintentional Harms of Mental Health Applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–17.
- Kim, J.; Rodriguez, V. J.; Yoo, D. W.; Chandrasekharan, E.; and Saha, K. 2026. PAIR-SAFE: A Paired-Agent Approach for Runtime Auditing and Refining AI-Mediated Mental Health Support. *arXiv preprint arXiv:2601.12754*.
- Kolla, M.; Salunkhe, S.; Chandrasekharan, E.; and Saha, K. 2024. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–8.
- Kumar, D.; AbuHashem, Y. A.; and Durumeric, Z. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 865–878.
- Li, J.; Zhu, Z.; Zhang, R.; and Lee, Y.-C. 2025a. Exploring the Effects of Chatbot Anthropomorphism and Human Empathy on Human Prosocial Behavior Toward Chatbots. *Proc. ACM Hum.-Comput. Interact.*, 9(7).
- Li, Y.; Yao, J.; Bunyi, J. B. S.; Frank, A. C.; Hwang, A.; and Liu, R. 2025b. CounselBench: A Large-Scale Expert Evaluation and Adversarial Benchmarking of Large Language Models in Mental Health Question Answering. *arXiv preprint arXiv:2506.08584*.
- Mahomed, Y.; et al. 2024. Auditing GPT's Content Moderation Guardrails: Can ChatGPT Write Your Favorite TV Show? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 660–686. Rio de Janeiro, Brazil.
- Markov, T.; Zhang, C.; Agarwal, S.; Eloundou, T.; Lee, T.; Adler, S.; Jiang, A.; and Weng, L. 2023. A Holistic Approach to Undesired Content Detection in the Real World. *arXiv:2208.03274*.
- Miller, W. R.; and Rollnick, S. 2013. *Motivational Interviewing: Helping People Change*. New York, NY, USA: Guilford Press, 3rd edition.
- Nie, J.; Shao, H. V.; Fan, Y.; Shao, Q.; You, H.; Preindl, M.; and Jiang, X. 2025. LLM-based Conversational AI Therapist for Daily Functioning Screening and Psychotherapeutic Intervention via Everyday Smart Devices. *ACM Transactions on Computing for Healthcare*. Just Accepted.
- NIH. 2024. Mental Illness. <https://www.nimh.nih.gov/health/statistics/mental-illness>. Accessed: 2025-10-11.
- OpenAI. 2024. Model Optimization Guide. <https://platform.openai.com/docs/guides/model-optimization>. Accessed: 2025-10-11.
- OpenAI. 2025. OpenAI omni-moderation-latest Model Documentation. <https://platform.openai.com/docs/models/omni-moderation-latest>.
- Pérez-Rosas, V.; Wu, X.; Resnicow, K.; and Mihalcea, R. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 926–935.
- Pham, K. T.; Nabizadeh, A.; and Selek, S. 2022. Artificial Intelligence and Chatbots in Psychiatry. *Psychiatric Quarterly*, 93(1): 249–253.
- Saha, K.; Jain, Y.; Liu, C.; Kaliappan, S.; and Karkar, R. 2025. AI vs. Humans for Online Support: Comparing the Language of Responses from LLMs and Online Communities of Alzheimer's Disease. *ACM Transactions on Computing for Healthcare*.
- Sharma, A.; Rushton, K.; Lin, I. W.; Nguyen, T.; and Althoff, T. 2024. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *Proc. CHI*.
- Shi, J. M.; Yoo, D. W.; Wang, K.; Rodriguez, V. J.; Karkar, R.; and Saha, K. 2026. Mapping Caregiver Needs to AI Chatbot Design: Strengths and Gaps in Mental Health Support for Alzheimer's and Dementia Caregivers. *ACM Transactions on Computing for Healthcare*.
- Song, M.; Kim, H.; Kim, J.; Shin, S.; and Son, S. 2025. Refusal Is Not an Option: Unlearning Safety Alignment of Large Language Models. In *34th USENIX Security Symposium (USENIX Security 25)*, 319–338.
- Wainberg, M. L.; Scorza, P.; Shultz, J. M.; Helpman, L.; Mootz, J. J.; Johnson, K. A.; Neria, Y.; Bradford, J.-M. E.; Oquendo, M. A.; and Arbuckle, M. R. 2017. Challenges and Opportunities in Global Mental Health: A Research-to-Practice Perspective. *Current Psychiatry Reports*, 19: 1–10.
- Yoo, D. W.; Shi, J. M.; Rodriguez, V. J.; and Saha, K. 2025. AI Chatbots for Mental Health: Values and Harms from Lived Experiences of Depression. *arXiv preprint arXiv:2504.18932*.
- Yuan, Y.; Zhang, J.; Aledavood, T.; Zhang, R.; and Saha, K. 2026. Mental Health Impacts of AI Companions: Triangulating Social Media Quasi-Experiments, User Perspectives, and Relational Theory. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*.
- Zhan, X.; Goyal, A.; Chen, Y.; Chandrasekharan, E.; and Saha, K. 2025. SLM-mod: Small language models surpass LLMs at content moderation. In *Proc. NAACL*.