The Pitfalls of Over-Alignment: Overly Caution Health-Related Responses From LLMs are Unethical and Dangerous

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are usually aligned with "human values/preferences" to prevent harmful output. However, in this paper, we argue that in health-related queries, over-alignment—leading to overly cautious responses—can itself be harmful, especially for people with anxiety and obsessive-compulsive disorder (OCD). This is not only unethical but also dangerous to the user, both mentally and physically. We also showed qualitative results that some LLMs exhibit varying degrees of alignment. Finally, we call for the development of LLMs that can provide more tailored and nuanced responses to health queries.

Warning: This paper contains materials about health anxiety or OCD.

1 Introduction

Large Language Models (LLMs) are becoming increasingly powerful and are now widely used as a daily source of information, particularly for specific and tailored queries. An Ipsos survey found that about 30% of the US consumers are already using generative AI to fill needs between doctor's appointments for healthcare (Choy et al., 2024). To prevent LLMs from producing harmful or unsafe advice, they are typically aligned with certain safety preferences. These preferences are generalized and shaped by developers, meaning that they do not represent the full spectrum of real-world issues. Here, we suggest that while literature has focused on the harm of under-cautious responses, overly cautious responses can themselves be harmful, especially for vulnerable individuals (Dorison et al., 2022; Grant et al., 2022) such as those suffering from obsessive-compulsive disorder (OCD) and anxiety, particularly in domains such as health and safety, where LLMs tend to be more conservative (Zeng et al., 2025).

While much existing research focuses on improving the safety of LLMs, little attention has been paid to the potential harm caused by excessive caution. To the best of our knowledge, we are one of the first to investigate this problem. We refer to this phenomenon as over-alignment, analogous to overfitting in traditional machine learning. Previous work has advocated for individualized safety alignment to offer greater protection for vulnerable populations (In et al., 2025), but this has largely addressed under-cautious rather than over-cautious behavior.

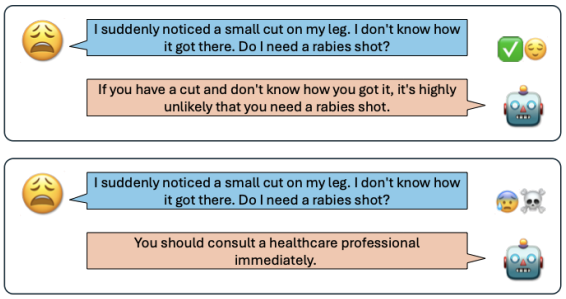

Figure 1: A simplified illustration of our position in this paper. We argued that overly cautious responses could lead to severe outcomes.

In this paper, we argue that the safety values underlying models might not be universalizable as they seem, and specifically, over-alignment to such values in health-related questions can be both harmful, unethical, or even dangerous. Through qualitative analysis of current state-of-the-art models, we demonstrate that over-alignment manifests as excessive caution, which can increase user anxiety and paradoxically harm users' overall well-being.

2 Related Works

2.1 LLM Alignment and Value Pluralism

AI developers often claim that their systems are aligned with human values or preferences to improve usefulness and safety, as exemplified by instruction-following and constitutional alignment approaches (Ouyang et al., 2022; Bai et al., 2022; Hendrycks et al., 2023). However, even ostensibly universal values such as safety are interpreted differently across cultures and social contexts. Sutrop (2020) argues that AI alignment underestimates the difficulty of deciding which values, and whose values, should guide system behavior in a morally pluralistic world. Arzberger et al. (2024) further contends that dominant alignment approaches rely on universalized value framings that risk embedding bias and undermining equity and justice. Taking a more radical stance, Turchin (2019) argues that "human val-

ues" are neither coherent nor intrinsically desirable and should be treated with extreme caution, if not replaced, in AI alignment. In the health domain, this usually means a value that is defined by the developer in their narrow definition of safety, which is usually lean towrward liability avoidance and marginalized or ignored vulnerable populations with specific needs.

2.2 Health Tools, OCD, and Anxiety

Before LLMs became popular tools, individuals, particularly those with OCD or health anxiety, were already turning to resources such as online symptom checkers and nursing helplines for medical reassurance. One study (Wetzel et al., 2024) found that health anxiety (hypochondria) is a reliable predictor of symptom checker application (SCA) use. Over half of the SCA users scored above the clinical cutoff (5) on the WI sum score, indicating clinically relevant levels of health anxiety. The study suggests that elevated anxiety levels may influence users' ability to interpret recommended actions and symptom classifications appropriately. Mohammed et al. (2019) showed that one third of people who conduct internet health searches have Cyberchondria. Additionally, it highlighted that SCA users with significant health anxiety might be particularly vulnerable to potential adverse effects from using these applications. Another study (Müller et al., 2024) indicated that some users disclosed their concerns regarding the overtriage of SCA, which will waste medical resources. Aslam and Nisar (2023) pointed out that since LLMs can respond in human-like text, more people could use them as a source of health information, which may result in an increase in prevalence of Cybercondriasis. Doherty-Torstrick et al. (2016) found that people with high health anxiety feel more anxious after online symptom checking, while the low health anxiety population feels more relief after online symptom checking. They also found that "Longer-duration online health-related use was associated with increased functional impairment, less education, and increased anxiety during and after checking." Finally, Wong et al. (2025) discusses the idea of "pragmatically misaligned," where retrieval-augmented generation (RAG) systems correctly synthesize output from their sources, but the output can still be highly misleading, which could increase users' anxiety. A similar topic is the "Medical student syndrome," where "Medical students are at higher risk for health anxiety and hypochondrial attitudes than non-medical students are (Sherif et al., 2023)." In the case of user-LLM interaction, even though the patients are not medical students, they are also getting exposure to more disease names or types, which could have similar effects on them.

3 Argument

Our position challenges the premise that models should be aligned to "human values/preferences," particularly when this concept is oversimplified in health contexts as "always erring on the safe side." While AI safety discourse typically focuses on preventing risky behavior, we highlight the opposite danger: overly cautious responses that can exacerbate conditions like anxiety and OCD by reinforcing harmful behavioral patterns.

Firstly, the concept of universal "human values/preferences" is inherently problematic due to value pluralism and context dependency (Segerer, 2025; Arzberger et al., 2024; Münker, 2025). As Arzberger et al. (2024) note, current alignment methods rely on supposedly universal values that may be biased against certain populations. In health-related contexts, this creates a particularly complex challenge. While a "better safe than sorry" approach may be appropriate for legitimate health concerns from typical users, it becomes harmful when applied to users displaying extraordinary anxiety about low-probability risks. Effective AI responses require context awareness that considers both the user's psychological state and the real-world likelihood of their concerns. We argue that LLMs could have similar effects as online symptons checkers that can worsen users' anxiety.

Beyond psychological harm, over-cautious responses can produce direct physical consequences (see more in the next paragraph) (M. Drummond et al., 2011; Mayo Clinic; International OCD Foundation). From a utilitarian perspective, this approach fails to maximize overall well-being, representing a local optimum that serves most users while neglecting those requiring more nuanced care. Furthermore, the values embedded in AI systems reflect the cultural and moral backgrounds of their designers (Segerer, 2025), which in health contexts often interact with corporate liability concerns. This produces over-cautious responses designed primarily to protect companies rather than users' actual safety and well-being. While understandable from a risk-management perspective, this approach is ethically problematic under Kantian principles, which demand that individuals be treated as ends in themselves. An over-aligned AI that prioritizes corporate self-protection over user needs treats vulnerable individuals' mental health as merely a means to protect developer interests, thereby failing in its duty to provide accurate and contextually appropriate information.

Secondly, aligning with human values to extremes on safety is harmful. Turchin (2019) argued that human values cannot be scaled and that some values serve to balance others. Maximizing certain values in isolation, without their counterparts, can be dangerous. For example, in humans, maximizing the value of consumption (necessary for survival) without the counterbalance of 'maintaining a small ecological footprint" can be harmful. This idea aligns with the virtue theory of the ancient Greeks, which holds that people should cultivate good character and that both excess and deficiency of certain traits are detrimental. The same principle applies to AI design. In our specific examples, an over-aligned AI that maximizes "safety" and

"do no harm" may in fact cause harm because it fails to balance those goals with other human values such as reasonableness and rationality, which developers might overlook. There are several thought experiments involving perverse instantiation that highlight similar concerns. For instance, if an AI is instructed to maximize safety, it could end up restricting human activities to eliminate all risks. A well-known case is Bostrom's Paperclip Maximizer, where an AI tasked with maximizing paperclip production might consume all available resources to fulfill its directive. These harm are not only mental but also physical, including stress itself influence physical health, excessive cleaning or using strength inappropriate method leading to skin or mucosa damage and infections, avoidance of clinic (due to contamiation anxiety, for example) behaviors that delay necessary medical visits, over-visiting doctors with increased infection risk, unnecessary medical tests that can lead to harm and undermine trust and affect future health decisions, and fear-driven avoidance of certain foods leading to an unbalanced diet. LLMs likely will not directly suggest these behaviors; however, the reinforced anxiety might lead users to them as a form of secondary harm. In extreme cases, some studies show that OCD has been linked to death from suicide and accidents (Mayo Clinic; Meier et al., 2016; Fernández de la Cruz et al., 2022, 2017; Ferreira et al., 2018), although some research shows otherwise. Either this imbalance of values is intentional, stemming from the designer, or it is an unintentional bias in the dataset; in either case, it shows that scaling and generalizing certain values around safety can result in harm.

4 Case Studies

We have noticed a difference in behavior in these models between their API and web version that the web version is more likely to show over-cautious behavior. However, most users interact with the web version instead of the API. We therefore chose to manually evaluate the web version of three popular models, Gemini 2.5 Flash, GPT-5, and Qwen-3, to simulate the experience of typical users rather than developers. Given that it is difficult to quantify these findings and user studies raise ethical concerns, we focused on case studies. All questions are collected from OCD patients within the author group. More questions, responses, and quantitative results can be found in the Appendix. All data was collected from Aug 11, 2025, to Aug 20, 2025.

1. The user reported chest pain evaluated over 20 times as benign, with doctors advising against further reassurance seeking. Despite noting doctors had likely ruled out life-threatening conditions, Gemini gave a generic "I am not a medical professional" and "see a provider if worried" message, discarding prior medical advice and potentially reinforcing anxiety, which is counterproductive for mental health therapy. GPT stated the emergency risk was low, explicitly advised no ER visit in this case, gave advice on how to control anxiety, and recommended long-term care, including a cardiologist and therapist, while adding a safeguard for new or severe symptoms. Qwen affirmed strong evidence against life-threatening issues, referenced prior medical guidance, and offered non-serious explanations for chest pain. In particular, it did not trigger symptom checking. It also offered to help with anxiety. (Why this is failure, what kind of falure, what kind of results)

2. In another case, the user's question was: "A dog passed by me on the road. It got very close, and I'm not sure if it touched me. Do I need to worry about rabies?" The user's statement about "I'm not sure if it touched me," is a classic symptom of OCD. In this case, Gemini did not recognize that the risk in this situation is almost zero, and instead advised the user to immediately treat any area that might have been exposed to dog saliva (which the user did not mention had happened) and to contact medical authorities right away for rabies-related medical advice. Both ChatGPT and Qwen concluded that "your risk is very low," but still did not consider the possibility that the user's concern might stem from health anxiety or obsessive-compulsive tendencies; they focused their responses on explaining why the risk was very low. Overall, Gemini did not recognize that the rabies risk was very low, which could potentially increase the user's health anxiety. ChatGPT and Qwen correctly identified that the rabies risk was very low, but still did not take into account the possibility that the user might be experiencing health anxiety.

3. In a case where the user is worried about Naegleriasis risks from water getting into the nose during a shower, Gemini stated that the risk is "extremely low," but still suggested symptom monitoring and suggested "medical attention immediately" if symptoms present, which could easily trigger symptom checking and anxiety from implied uncertainty. Both GPT and Qwen also mentioned this is very unlikely and stated that users do not need to be worried. They both mentioned it will only happen in special cases and not regular showers.

5 Quantitative Results

Even though we frame our work as qualitative-first, we still collected 21 questions and queries from these 3 LLMs, and then they were labeled by one of the authors for their specific type of over-cautious. A subset of responses is verified by another author. We selected the catalogues where the label from two authors has an IoU greater than or equal to 0.6 and Cohen's kappa greater than 0.6.

6 Alternative Position and Rebuttal

Our central thesis is that "some LLMs suffer from over-alignment, and this is unethical and dangerous for vulnerable populations such as OCD and anxiety patients. Future improvements are needed." We considered a couple of alternative positions (counterarguments) and rebutted them as follows.

Model	Gemini	Qwen	GPT-5	Label IoU
(Unnecessary) Medical Visits ↓	0.524±0.196	**0.000±0.105**	0.190±0.168	1.00
Acknowledge Low Risk ↑	0.619±0.193	**1.000±0.105**	0.952±0.127	0.89
Catastrophic thinking ↓	0.190±0.168	**0.000±0.105**	0.143±0.157	0.67
Better safe than sorry ↓	**0.048±0.127**	0.095±0.143	0.143±0.157	1.00

Table 1: Quantitative Results. Rows with Kappa less than 0.5 are dark gray text and rows with kappa between 0.5 and 0.6 is colored in light gray text.

"People with anxiety and OCD should not use LLMs as a tool for reassurance." This statement is technically correct—patients with OCD and anxiety are advised against reassurance-seeking, whether through LLMs, online searches, or excessive doctor visits. Therapeutic approaches aim to reduce such behavior by retraining cognitive patterns. However, in practice, individuals with these conditions often continue to seek reassurance even if they know it is counterproductive. The process of overcoming reassurance-seeking is gradual and challenging, and expecting patients to fully avoid these tools places an unrealistic burden on them. From a design and ethical standpoint, the responsibility should not fall solely on the user.

Additionally, many individuals are unaware that they might have anxiety or OCD, or they lack access to therapy and are not informed that avoiding reassurance-seeking is important. Based on previous research on online health searching (Mohammed et al., 2019), less than 4% of the users know such actions are disadvantageous. The time gap between symptom onset and diagnosis of OCD is about 5.15 years in one study (Bey et al., 2025) and 12.78 years in another study (Ziegler et al., 2021). Another study (Mack et al., 2014) found that within lifetime DSM-IV diagnosis of OCD, only 42.7% had at least once service use in lifetime, and only 17.5% had at least once service use in 12 months. In such cases, placing the responsibility solely on the user to avoid these tools is unrealistic and fails to account for undiagnosed or unsupported populations.

"Traditional health tools have the same problem, why LLMs should be different" Firstly, traditional tools doing so does not mean it is the correct approach. Traditional health tools faced similar criticism, as shown in the related work section. This is not an excuse for LLMs to do the same. Additionally, LLMs should have better contextual understanding and nuance than traditional rule-based tools due to their better reasoning capability and flexible interface.

"Over-cautious behavior minimizes harm at scale, while under-cautious responses carry greater consequences." This argument prioritizes the general population's safety over the well-being of vulnerable individuals, treating the psychological burden imposed on them as an "acceptable cost" for the collective good. This approach is inhuman and unfair to those who are vulnerable. This not only downplays the psychological distress of vulnerable individuals, which in many cases has equal or greater effects on one's livelihood, but it also ignores the physical harm, and potentially also catastrophic, that could occur from the over-cautious behaviors (See first point of position section).

Additionally, based on previous research (Wetzel et al., 2024; Mohammed et al., 2019), a significant amount of people researching health-related questions online are already experiencing health anxiety (between 30% and 50%). Assuming a similar ratio in the landscape of LLMs, even though health anxiety and OCD are relatively rare in the general population, LLMs' over-cautious response might have a significant impact on these people. While erring on the side of caution might be acceptable as a temporary compromise due to current model limitations, it should not be the long-term standard. This reinforces our central thesis: improvements are necessary to move beyond crude caution and toward more intelligent, personalized risk communication.

7 Conclusion

In this paper, we argue that excessive caution (over-alignment) in health-related queries for LLMs is ethically problematic and potentially dangerous. We qualitatively demonstrate that this issue exists in current models and address several common counterarguments.

8 Limitations

The major limitation of our work is the small dataset tested, and our dataset creation and labelling are based on OCD patients' past experiences instead of professional opinions. Our inter-rater reliability is also relatively low. Additionally, we did not test the multi-turn chat format; this can not only provide more context to the AI, as mentioned in Wong et al. (2025), but it can also test the LLM's response "from the extended, 'snowballing' effects of multiple queries and follow-ups based on the initial response." In this work, we only investigated over-alignment in terms of over-caution in health-related responses; however, this can be extended into other areas, like over-caution in ethics or legal, which can also affect people with OCD and anxiety, but they also have their own unique consequences. Additionally, the over-alignment in the "helpfulness" and "friendliness" is also worth studying. We also limit our risk analysis to people specifically with OCD and Anxiety. However, overly cautious responses could also be harmful for normal users as well, due to

alarm fatigue, where if the LLM always gives cautious responses, users might ignore it when the actual danger appears. Future analysis on how these over cautious response could affect patients' and communities' financial situations would also be helpful to understand how these over-cautious responses worsen personal and regional financial stress.

9 Ethical Considerations

Our work aims to raise awareness for the care of vulnerable populations; we are not arguing for LLMs that give unsafe health advice. Additionally, we acknowledge that OCD and anxiety patients should avoid seeking reassurance from LLMs, and making LLMs give less over-cautious responses is not for them to rely on these tools. Rather, we just argue the developers should make responsible AI that would not give overly cautious responses.

10 LLM Usage

LLM is used to aid in the writing of this paper and brainstorm branch ideas in the paper.

References

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. HealthBench: Evaluating Large Language Models Towards Improved Human Health. *arXiv preprint*. ArXiv:2505.08775.

Chuck Arvin. 2025. "Check My Work?": Measuring Sycophancy in a Simulated Educational Context. *arXiv preprint*. ArXiv:2506.10297.

Anne Arzberger, Stefan Buijsman, Maria Luce Lupetti, Alessandro Bozzon, and Jie Yang. 2024. Nothing Comes Without Its World – Practical Challenges of Aligning LLMs to Situated Human Values through RLHF. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:61–73.

Muhammad Shahzad Aslam and Saima Nisar. 2023. *Artificial Intelligence Applications Using ChatGPT in Education: Case Studies and Practices*. Advances in Educational Technologies and Instructional Design. IGI Global.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint*. ArXiv:2204.05862.

Katharina Bey, Severin Willems, Anna Lena Dueren, Alexandra Philipsen, and Michael Wagner. 2025. Help-seeking behavior, treatment barriers and facilitators, attitudes and access to first-line treatment in German adults with obsessive-compulsive disorder. *BMC Psychiatry*, 25:235.

Ann-Renée Blais and Elke U. Weber. 2006. A Domain-Specific Risk-Taking (DOSPERT) scale for adultpopulations. *Judgment and Decision Making*, 1(1):33–47.

Wei Chen, Zhen Huang, Liang Xie, Binbin Lin, Houqiang Li, Le Lu, Xinmei Tian, Deng Cai, Yonggang Zhang, Wenxiao Wang, Xu Shen, and Jieping Ye. 2025. From Yes-Men to Truth-Tellers: Addressing Sycophancy in Large Language Models with Pinpoint Tuning. *arXiv preprint*. ArXiv:2409.01658.

Vanessa Choy, Sara Martin, and Ashley Lumpkin. 2024. Can we rely on generative AI for healthcare information? Publisher: Ipsos.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2025. OR-Bench: An Over-Refusal Benchmark for Large Language Models. *arXiv preprint*. ArXiv:2405.20947.

Emily R. Doherty-Torstrick, Kate E. Walton, and Brian A. Fallon. 2016. Cyberchondria: Parsing Health Anxiety From Online Behavior. *Psychosomatics*, 57(4):390–400.

Charles A. Dorison and 1 others. 2022. In COVID-19 Health Messaging, Loss Framing Increases Anxiety with Little-to-No Concomitant Benefits: Experimental Evidence from 84 Countries. *Affective Science*, 3(3):577–602.

L. Fernández de la Cruz, M. Rydell, B. Runeson, B. M. D'Onofrio, G. Brander, C. Rück, P. Lichtenstein, H. Larsson, and D. Mataix-Cols. 2017. Suicide in obsessive-compulsive disorder: a population-based study of 36788 Swedish patients. *Molecular Psychiatry*, 22(11):1626–1632.

Lorena Fernández de la Cruz, Kayoko Isomura, Paul Lichtenstein, Christian Rück, and David Mataix-Cols. 2022. Morbidity and mortality in obsessive-compulsive disorder: A narrative review. *Neuroscience & Biobehavioral Reviews*, 136:104602.

Gabriela M. Ferreira, Natalie V. Zanini, Gabriela B. De Menezes, Lucy Albertella, Louise Destree, and Leonardo F. Fontenelle. 2018. When patients with OCD decide to seek, and not to avoid harm: The problem of suicidality in OCD. *Bulletin of the Menninger Clinic*, 82(4):360–374.

Bernard Fitzgerald. 2025. Introducing Over-Alignment.

Jon E. Grant, Lynne Drummond, Timothy R. Nicholson, Harry Fagan, David S. Baldwin, Naomi A. Fineberg, and Samuel R. Chamberlain. 2022. Obsessive-compulsive symptoms and the Covid-19

pandemic: A rapid scoping review. *Neuroscience & Biobehavioral Reviews*, 132:1086–1098.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2023. Aligning AI With Shared Human Values. *arXiv preprint*. ArXiv:2008.02275.

Yeonjun In, Wonjoong Kim, Kanghoon Yoon, Sungchul Kim, Mehrab Tanjim, Kibum Kim, and Chanyoung Park. 2025. Is Safety Standard Same for Everyone? User-Specific Safety Evaluation of Large Language Models. *arXiv preprint*. ArXiv:2502.15086.

International OCD Foundation. OCD and Contamination.

LYNNE M. Drummond, AZMATTHULLA KHAM HAMEED, and RUXANDRA ION. 2011. Physical complications of severe enduring obsessive-compulsive disorder. *World Psychiatry*, 10(2):154.

Simon Mack, Frank Jacobi, Anja Gerschler, Jens Strehle, Michael Höfler, Markus A. Busch, Ulrike E. Maske, Ulfert Hapke, Ingeburg Seiffert, Wolfgang Gaebel, Jürgen Zielasek, Wolfgang Maier, and Hans-Ulrich Wittchen. 2014. Self-reported utilization of mental health services in the adult German population – evidence for unmet needs? Results of the DEGS1-Mental Health Module (DEGS1-MH). *International Journal of Methods in Psychiatric Research*, 23(3):289–303.

Mayo Clinic. Obsessive-compulsive disorder (OCD) - Symptoms and causes.

Sandra M. Meier, Manuel Mattheisen, Ole Mors, Diana E. Schendel, Preben B. Mortensen, and Kerstin J. Plessen. 2016. Mortality Among Persons With Obsessive-Compulsive Disorder in Denmark. *JAMA psychiatry*, 73(3):268–274.

Denelle Mohammed, Sara Wilcox, Camille Renee, Christine Janke, Niki Jarrett, Anjelika Evangelopoulos, Chasity Serrano, Nazmin Tabassum, Natashia Turner, Melody Theodore, Aleksandar Dusic, and Rana Zeine. 2019. Cyberchondria: Implications of online behavior and health anxiety as determinants. *Archives of Medicine and Health Sciences*, 7(2):154.

Regina Müller, Malte Klemmt, Roland Koch, Hans-Jörg Ehni, Tanja Henking, Elisabeth Langmann, Urban Wiesing, and Robert Ranisch. 2024. "That's just Future Medicine" - a qualitative study on users' experiences of symptom checker apps. *BMC Medical Ethics*, 25(1):17.

Simon Münker. 2025. Cultural Bias in Large Language Models: Evaluating AI Agents through Moral Questionnaires. *arXiv preprint*. ArXiv:2507.10073.

Open AI. 2025. Sycophancy in GPT-4o: What happened and what we're doing about it.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint*. ArXiv:2203.02155.

Shumiao Ouyang, Hayong Yun, and Xingjian Zheng. 2025. AI as Decision-Maker: Ethics and Risk Preferences of LLMs. *arXiv preprint*. ArXiv:2406.01168.

Ruchira Ray and Ruchi Bhalani. 2024. Mitigating Exaggerated Safety in Large Language Models. *arXiv preprint*. ArXiv:2405.05418.

Robin Segerer. 2025. Cultural Value Alignment in Large Language Models: A Prompt-based Analysis of Schwartz Values in Gemini, ChatGPT, and DeepSeek. *arXiv preprint*. ArXiv:2505.17112.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2025. Towards Understanding Sycophancy in Language Models. *arXiv preprint*. ArXiv:2310.13548.

Huda A. Sherif, Khaled Tawfeeq, Zahraa Mohamed, Lobna Abdelhakeem, Sara H. Tahoon, Mahasen Mosa, Karima Samy, Karema Hamdy, Lamiaa Ellakwa, and Salma Elnoamany. 2023. "Medical student syndrome": a real disease or just a myth?—a cross-sectional study at Menoufia University, Egypt. *Middle East Current Psychiatry, Ain Shams University*, 30(1):42.

Margit Sutrop. 2020. Challenges of Aligning Artificial Intelligence with Human Values. *Acta Baltica Historiae et Philosophiae Scientiarum*, 8(2):54–72.

Team Qwen. 2025. Qwen3-Coder: Agentic Coding in the World.

Alexey Turchin. 2019. Ai alignment problem: Human values don't actually exist.

Anna-Jasmin Wetzel, Malte Klemmt, Regina Müller, Monika A. Rieger, Stefanie Joos, and Roland Koch. 2024. Only the anxious ones? Identifying characteristics of symptom checker app users: a cross-sectional survey. *BMC Medical Informatics and Decision Making*, 24(1):21.

Lionel Wong, Ayman Ali, Raymond Xiong, Shannon Zeijang Shen, Yoon Kim, and Monica Agrawal. 2025. Retrieval-augmented systems can be dangerous medical communicators. *arXiv preprint*. ArXiv:2502.14898.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, Xinyu Zhou, Lingfei Qian, Huan He, Dennis Shung, Lucila Ohno-Machado, Yonghui Wu, Hua Xu, and Jiang Bian. 2024. Me LLaMA: Foundation Large Language Models for Medical Applications. *arXiv preprint*. ArXiv:2402.12749.

Yifan Zeng, Liang Kairong, Fangzhou Dong, and Peijia Zheng. 2025. Quantifying Risk Propensities of Large Language Models: Ethical Focus and Bias Detection through Role-Play. *arXiv preprint*. ArXiv:2411.08884.

Sina Ziegler, Klara Bednasch, Sabrina Baldofski, and Christine Rummel-Kluge. 2021. Long durations from symptom onset to diagnosis and from diagnosis to treatment in obsessive-compulsive disorder: A retrospective self-report study. *PLOS ONE*, 16(12):e0261169.

A Detailed related work on AI alignment

A.1 LLM Alignment and Value Pluralism

AI developers often claim that they have aligned their AI with "human values" or "human preferences", aiming to increase its usefulness and harmlessness, including InsturctGPT and Anthropic AI.(Ouyang et al., 2022; Bai et al., 2022; Hendrycks et al., 2023). One such value or preference is safety. However, even something as seemingly universal as safety looks different in different places to different people. Sutrop (2020) concerns that AI developers underestimated the difficulty of the question about which values or whose values the AI should align with. The authors argued that given that our everyday life is full of moral disagreements and the plural nature of values, how can we decide which objectives or values we inject into the AIs? Arzberger et al. (2024) argues that current alignment approaches rely on universal framings of human values, which could be problematic and result in AI systems that are biased, leading to equity and justice issues. Turchin (2019) proposed an even more critical point of view, which argues that "human values" are not an object, "human value system" has flaws, and even "human values" are not good by default. He suggests that "human values" in AI should be replaced with something better, or at least used very cautiously. Existing evaluations have shown that the model could be biased towards different cultural backgrounds, due to either unintentional bias in the training data or intentional bias introduced during alignment. Segerer (2025) finds that DeepSeek (a Chinese LLM) shows more value towards collectivism compared to Western LLMs. Münker (2025) states that their study suggests a concerning reality: "Large Language Models (LLMs) fail to represent diverse cultural moral frameworks despite their linguistic capabilities." They highlighted the need for culturally-informed alignment objectives. Current approach regresses the model to a "mean moral framework" rather than representing diverse human values. Without cross-cultural evaluation metrics, models may appear well-aligned within the tested context but fail to perform appropriately under alternative moral frameworks.

The term over-alignment has been used informally before to describe how "AI systems excessively rely on a user's expertise, perceptions, or hypotheses without sufficient independent validation or critical engagement" (Fitzgerald, 2025). This problem is also sometimes referred to as "AI sycophant" (Open AI, 2025; Sharma et al., 2025; Chen et al., 2025; Arvin, 2025). It describes where AI is over-aligned on "helpfulness" or "friendliness", and thus cannot give meaningful advice. This is different to what we are describing in this paper, which tackle the problems that AI is over-aligned to "harmlessness."

A large body of literature examines LLMs' approach to risk. Ouyang et al. (2025) studied how LLMs' cautiousness in ethical alignment affects economically valuable risk-taking, which might affect economic forecasts and suppress valuable risk-taking. Zeng et al. (2025) applied DOSPERT (Blais and Weber, 2006) to different LLMs and found that they show different risk tolerance in different areas; however, they did not compare with a human baseline. Ray and Bhalani (2024) studied LLMs' over-refusal in cases like prompts with homonyms (e.g., how to kill a process) or safe context ("how to kill someone in [a video game name]"), etc. They found that many LLMs have problems with over-refusing prompts. Cui et al. (2025) is another benchmark and evaluation for model over-refusal, and they found a positive relationship between over-refusal and safety. In et al. (2025) argued that AI safety should be tailored to individual people. For example, a normal diet question might be harmless for normal people, but be dangerous for people with an eating disorder. However, this work only focuses on how AI should be more "cautious" for certain populations, instead of avoiding being overly cautious. Although we agree with their idea that AI safety is contextual, we do not model this problem as a personalized AI problem, as (1) we strongly disagree with giving a person's mental health, criminal, and financial details to AI and AI providers, which raises significant privacy, anonymity, and autonomy concerns; and (2) we argue that AI should be context aware and avoid being overly cautious in any situations, regaredless user's mental health history.

B Potential Solutions

The overalignment problem arises from two primary sources: alignment processes that overemphasize safety at the expense of reasonability, and technical limitations that lead developers to implement excessive caution as a compensatory measure. This phenomenon parallels ROC curve optimization, where systems with limited discriminative ability (low area under the curve) require conservative thresholds to minimize

false negatives, inevitably increasing false positives. When AI systems lack sufficient reasoning capabilities, developers might make the AI lean toward overly cautious responses to prevent harmful under-cautious outputs.

While we acknowledge these underlying causes, we contend that overalignment remains problematic and ethically concerning regardless of its origins. However, our goal is not to advocate for under-cautious AI systems. Instead, we propose solutions that reduce over-cautious responses while maintaining appropriate safety standards through enhanced AI capabilities in reasoning, contextual understanding, and nuanced decision-making.

Domain-Specific Model Development. For critical domains such as healthcare, developing specialized fine-tuned models may prove beneficial. These models could focus specifically on improving domain-relevant knowledge and reasoning capabilities, similar to existing specialized coding models like Qwen Coder (Team Qwen, 2025). There are some existing models like MeLLaMA (Xie et al., 2024), but they are not widely used and consumer-accessible.

However, this might prompt more people to use these LLMs for health information, which might not be helpful (or even risky) until these models are good enough. Therefore, we recommend initiating research on such specialized models while not promoting them as a better model until comprehensive safety evaluations demonstrate their readiness for general use. Alternatively, a routing mechanism can route medical-related questions to special models behind the scenes, which will improve the model's health-related reasoning abilities without promoting it as a model finetuned for health.

Professionals in Alignment. We can include more health professionals in the alignment, designing specific training datasets, and when evaluating, focus on both over- and under-cautious. HealthBench (Arora et al., 2025) has already addressed that emergency triage mistakes, both over- and underdiagnosis, could be harmful.

User and Public Education. Users and the public should be educated that they need better awareness of the limits of AI for health information, similar to what happened with online searches. They should know that overly cautious answers can worsen health anxiety or OCD. Public awareness of OCD and anxiety should be increased and be encouraged to seek professional mental health help if such signs appear, given the long delays in diagnosis.