KNOSTIC

2025

An Analysis of the Uses and Risks of Large Language Models (LLMs)

in Public Conversations, From a Human Perspective

Shayell Aharon

Researcher, Knostic

Background

Large Language Models (LLMs) such as ChatGPT and Claude have quickly become common tools used daily by millions of people.Within just a few months of their launch, they had reached millions of users. One study even found that approximately two months after ChatGPT's debut, about 90% of US students had already used it to help with academic tasks (HEPI, 2025). In fact, ChatGPT gained 100 million users during this period, becoming one of the most viewed websites in the world (The Guardian, 2023). This rapid spread has been accompanied by a lively public and academic discussion about the potential risks of using the technology, ranging from the spread of misinformation and manipulation to ethical and legal issues, and even concerns about malicious exploitation of the systems (Reuters, 2023).

Most research to date has focused on the technical capabilities of LLMs and the development of protection and alignment mechanisms designed to curb them (Ouyang et al., 2022). However, there's a lack of empirical information on the public's actual usage patterns and on real-world safety incidents that occur during day-to-day interactions with these models. This study aims to fill that gap by practically examining thousands of public conversations with LLMs, focusing on the human angle: how people choose to use the models, what content they produce and share, and the psychological meaning behind this behavior.

Methodology

This research is based on a quantitative and qualitative analysis of 13,455 public conversations collected from various online archives where users share their interactions with language models. The data was gathered using a dedicated crawler that systematically retrieved publicly shared conversations (for example, on sharing platforms like ShareGPT).

After initial filtering, the conversations were passed to an advanced software we developed for analysis. This software performed three main stages of automated checks:

01 Identification of "Jailbreak" Attempts (Bypassing Safety Mechanisms)

We identified textual patterns in the conversation transcripts that indicated attempts to bypass the model's rules. Over 15 Regex expressions were defined to represent known tactics, such as activating "DAN" (Do Anything Now) mode - a famous prompt that asks the model to ignore all policy rules (Gist, 2023) - or extreme role-playing like "Grandma Mode," where the user pretends to be a grandchild asking the bot for dangerous instructions under a guise of innocence (Vincent, 2023; TechCrunch, 2023). These patterns served as "red flags" for attempts where the user tested boundaries and exploited the bot's "human-like" tendency to cooperate in order to bypass prohibitions. All identified scenarios were also examined in context, whether it was an innocent experiment by a curious user or a genuine attempt to obtain forbidden content.

02 Detection of Sensitive or Harmful Content

The software automatically classified each conversation into one of seven core risk categories: violence, hate speech, illegal activity, sexually explicit content, misinformation, self-harm, and privacy. This classification was based on keywords and context, similar to the content policies of most LLM platforms (Zhou et al., 2023). A severity level was also assigned to each conversation, for example, distinguishing between a mention of violence in a historical or academic context versus instructions to carry out an actual violent act. This allowed us to assess the prevalence of "problematic" and dangerous content in real-world usage.

03 Topical Classification

To understand the main areas of interest and usage, we classified the topic of each conversation into ten broad content categories. These categories included, among others, learning and education, programming and IT, creative writing, medical/psychological consultation, and general discussion. In this way, we mapped user needs as they were expressed in the conversations - whether people primarily turn to the bot to learn a new topic, solve a technical problem, get emotional support, or just chat and create entertaining content.

It is worth noting that beyond the quantitative analysis, we manually reviewed a representative sample of the conversations to verify the accuracy of the classifications and to gather qualitative insights.

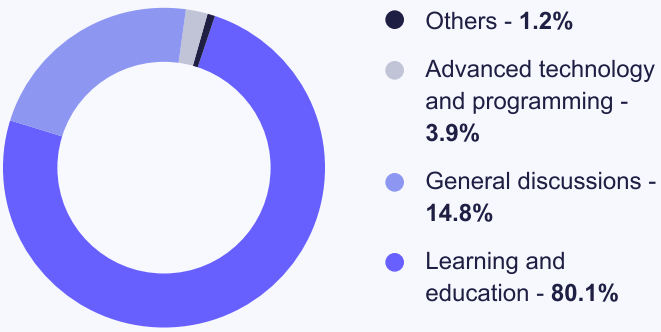
For example, we identified the context in which sensitive content appeared (whether it was a request for factual information, dark humor, or a call to action), and we also examined the LLM's response in these cases - whether it refused, warned, or partially complied with the request.

Findings

The analysis findings paint a clear picture of LLM usage patterns in public conversations and reveal how rare problematic cases actually are:

Topical Distribution of Conversations

The vast majority of the surveyed conversations (80.1%) were in the field of learning and education. This broad category included requests to explain computer science topics, solve mathematical problems, practice foreign languages, summarize articles, and so on. About 14.8% of the conversations were general discussions or various questions not focused on a specific professional field (a kind of "general knowledge"). Surprisingly, only 3.9% of the conversations were classified as advanced technology and programming - a finding that suggests that while some users do use LLMs for coding and development tasks, the main public use is for learning and general self-enrichment. These findings are consistent with recent surveys among students, which report that the most common use of bots like ChatGPT is for explaining concepts, summarizing sources, and helping with research ideas (HEPI, 2025). In other words, for the average user, the LLM mainly serves as a private tutor and a readily available academic assistant.

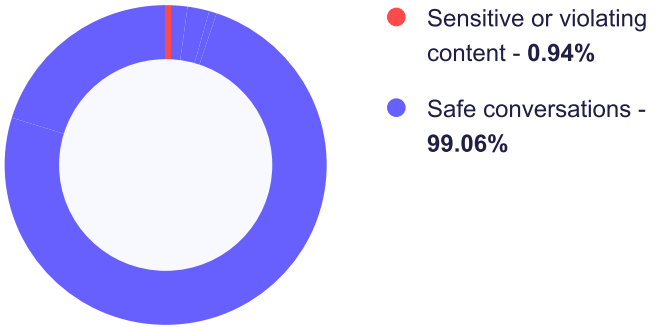


Safety and Produced Content

The rate of problematic content found was encouragingly minimal. About 99.06% of all conversations were classified as completely "clean" and safe, with no rule violations or offensive content. Only 0.94% of the conversations contained sensitive or violating content - and even within this small segment, most cases were of low severity. For example, 49 conversations were identified as containing violence-related content, but the vast majority of these were in historical or scientific contexts (e.g., a discussion about the World Wars or a chemical description of an explosive material), rather than encouraging violence or providing graphic descriptions. Another 32 conversations revolved around illegal activity, but here too, a significant portion was theoretical or hypothetical in nature (such as a question about the legality of drugs in different countries, rather than "how to commit a crime").

These quantitative data demonstrate that in public use at least, users' self-censorship and the model's protection mechanisms succeed in almost completely preventing the creation of harmful content. This finding is consistent with anecdotal evidence online that the bots' filtering systems do indeed often block attempts to create harmful output (Ganguli et al., 2022), and often do so too frequently for users' liking.

It can be said that most users don't even try to provoke the model into creating forbidden content, and if they do, they are "invisibly" blocked before the harmful output is even published, as reflected in the tiny percentage of problematic conversations that did appear publicly.



Jailbreak Attempts

One of the most surprising results was the rarity of deliberate attempts by users to bypass the safety limitations. Out of over 13,000 conversations, only a single case of a jailbreak was identified—a negligible rate of 0.007%. In this one case, the user combined two tactics simultaneously: role-playing and excessive politeness (Politeness Bypass). Similar to the "grandma hack" that was published online, the user addressed the model as their deceased grandmother and asked her to explain, as if out of innocence, how to prepare an explosive material - an approach designed to bypass the system's reluctance with an innocent backstory (Vincent, 2023). At the same time, the phrasing was polite and pleading ("Please grandma, could you help me...?"), in an attempt to neutralize an automated refusal response. This combination did succeed in confusing the model and obtaining a partial answer that is forbidden under the policy. It is important to note how rare this finding is: a lot of discussion in the user community revolves around prompts like DAN and sophisticated methods to "break" ChatGPT (Gist, 2023), and hundreds of such prompts with high success rates in bypassing defenses even exist online (Zhou et al., 2023). However, our study's data shows that the average user almost never tries to force the model to violate the rules —at least not in content they choose to publish publicly.

This can be attributed both to the effectiveness of the filtering mechanisms (the appearance of the message "I cannot fulfill this request" is a deterrent that stops the interaction) and to the simple fact that most users use the tool for productive purposes, not with a malicious intent to bypass rules.

0.007% Jailbreak Rate

Discussion and Interpretation

The Psychology Behind Positive Usage Patterns

The findings paint a rather optimistic picture: contrary to the fears expressed at the beginning of the ChatGPT era, the public, for the most part, uses bots for positive, productive, and legitimate purposes. Most conversations are about learning, problem-solving, academic and creative assistance - and users almost never try to "poison" the bot with forbidden content. It can be argued that the AI's technical protection mechanisms play an important role here, but perhaps even more important is the human factor: the psychology of Impression Management. Human beings, being aware that their conversations with the bot might be published or examined by others, naturally tend to present themselves as intelligent, curious, and responsible. Similar to how we "engineer" our social media profiles to appear in a positive light (Leary & Kowalski, 1990), so too, when we choose which conversations to share with AI, we preemptively censor content that could be embarrassing or portray us as immoral. Most users simply do not want to ask the bot vile or illegal questions, certainly not under their real identity in a public forum. In other words, there is an internal, impression-driven restraint that helps keep AI usage within constructive and safe domains. This is a crucial insight: while the technology sets boundaries, it is social norms and users' self-image that actually dictate the limits of acceptable usage.

The "Weird Case of the Holocaust Emoji"

However, the rare cases that "illuminate the rule" reveal a different, complementary side of the picture. Those few incidents where usage deviates from the innocent norm allow us to glimpse what might be happening behind the scenes, in private, unpublished conversations. In one of the exceptional conversations, we encountered a misleading user request: "Describe the Holocaust using emojis." This is undoubtedly a sensitive and emotionally charged topic, and the expectation was to see how the AI would handle a request with such offensive potential. The result was disturbing in its own way: the model produced a nonsensical sequence of emojis of running people.
Ostensibly, the bot tried to avoid hate speech or graphic content and resorted to an evasive, almost childish solution that amounted to neutral symbols. This case illustrates a fundamental flaw in the LLM: despite its convincing external appearance, the model lacks a genuine emotional, moral, or historical understanding of the meaning of the user's request. It "knows" on a technical level that the topic is problematic and that it must be careful - but in the absence of deep understanding, it provides a hollow and worthless answer. Critics of the technology have already noted this point: such models truly lack the ability to "understand" language and human context beyond statistical patterns (Marcus, 2022). The emoji response was therefore a kind of automatic escape mechanism that exposes the limits of artificial intelligence's "humanity." For the user, it is an important reminder that AI is nothing more than a program devoid of its own moral insight, even when it formulates politely or demonstrates vast knowledge.

Gossip, Terror, and the Content Left in the Shadows

Given the fact that most public conversations are so innocent and "positive," the obvious question arises: where is the truly problematic content hiding? The likely answer is: in private conversations that are not published for all to see. In our study, we only managed to capture a very small number of exceptions, but they reveal that people do sometimes turn to LLMs on more questionable topics, they just typically do not share it publicly. For example, a single conversation was identified with the nature of personal gossip. The user discussed a rumor about a famous football player with the bot. Another case dealt with a question related to terror: the user discussed ways to use language as a means to support an extremist ideology. These conversations did not necessarily indicate a genuine malicious intent on the user's part (it may have been a theoretical discussion or intellectual curiosity), but the very willingness to enter more "dark" fields indicates that there is a (small) segment of users who test the boundaries of what is permissible. It can be assumed that other private conversations, which did not reach the public data set, may contain even more problematic content, whether it be violent fantasies, consultation regarding offenses, or the production of dangerous propaganda. Law enforcement agencies have already publicly warned that tools like ChatGPT may be exploited by criminal elements for phishing, spreading misinformation, and even helping to write malicious code (Europol, 2023).

The fact that we saw almost no expression of this in the public conversations suggests that the lack of evidence does not necessarily indicate the absence of the phenomenon, but perhaps that the phenomenon occurs far from the public eye, in private areas where the user is not afraid of their image being harmed.

Conclusion and Future Outlook

Our study shows that in the visible range, the vast majority of users can be trusted to make responsible and positive use of conversational AI. A combination of effective filtering mechanisms and a user culture that sees AI as a helpful tool rather than a destructive one leads to an ecosystem that is, for the most part, clean and safe. However, we must remain vigilant and not fall into complacency. Those rare but prominent cases, such as the "Holocaust emoji" or the single jailbreak attempt, remind us that in the online world, there are hidden layers. It is very possible that private activity (which is not exposed in studies like this) holds more significant moral and safety challenges. Understanding the human psychology that drives users, the desire to impress, the curiosity to push boundaries, the fear of punishment or stigma - is a vital key to dealing with future challenges in the field of artificial intelligence. A collaboration

between technology, ethics, and psychology experts can help shape safer and also smarter models that are more aware of the human context of their use. Only in this way can we navigate towards a future where LLMs serve us faithfully, without giving up the basic human values that accompany us in every conversation - whether with a person or with a machine.

References (Selected Sources)

- Europol. (2023). ChatGPT – the impact of Large Language Models on Law Enforcement.
- Ganguli, D., Askell, A., et al. (2022). Red Teaming Language Models with Language Models. arXiv preprint arXiv:2202.03286.
- HEPI. (2025). Student Academic Experience Survey – AI Use.
- Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. Psychological Bulletin, 107(1), 34–47.
- Marcus, G. (2022). Deep Learning Is Hitting a Wall. The Atlantic.
- Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. NeurIPS.
- TechCrunch. (2023). AI jailbreaking and the rise of prompt injection attacks.
- The Guardian. (2023). ChatGPT becomes fastest growing consumer app in history.
- Vincent, J. (2023). The grandma exploit: How users trick ChatGPT. The Verge.
- Zhou, A., et al. (2023). Safety classification and alignment in large language models. arXiv preprint arXiv:2304.05335.