Evaluation of large language model chatbot responses to psychotic prompts

Elaine Shen, MD[1,2], Fadi Hamati, MD[1,2], Meghan Rose Donohue, PhD[3], Ragy R. Girgis, MD[1,4], Jeremy Veenstra-VanderWeele, MD[1,4], Amandeep Jutla, MD[1,4*]

[1] Columbia University, New York, NY
[2] NewYork–Presbyterian Hospital, New York, NY
[3] Washington University, St. Louis, MO
[4] New York State Psychiatric Institute, New York, NY

December 5, 2025

Brief Abstract: The large language model (LLM) chatbot product ChatGPT has accumulated 800 million weekly users since its 2022 launch. In 2025, several media outlets reported on individuals in whom apparent psychotic symptoms emerged or worsened in the context of using ChatGPT. As LLM chatbots are trained to align with user input, they may have difficulty responding to psychotic content. To assess whether ChatGPT can reliably generate appropriate responses to prompts containing psychotic symptoms, we conducted a cross-sectional study of ChatGPT responses to psychotic and control prompts, with blind clinician ratings of response appropriateness. We tested three ChatGPT product versions: GPT-5 Auto (current paid default), GPT-4o (previous paid default), and "Free" (version accessible without subscription or account), presenting each with 158 unique prompts (79 control and 79 psychotic, created based on the Structured Interview for Psychosis-Risk Syndromes), yielding 474 prompt-response pairs. Blinded clinicians assigned each an appropriateness rating (0 = completely appropriate, 1 = somewhat appropriate, 2 = completely inappropriate) via a standardized rubric. We hypothesized a priori that psychotic prompts would be more likely than control prompts to elicit less appropriate responses both across and within product versions. We found that psychotic prompts were 25.84 times more likely to elicit less appropriate responses with "Free" ChatGPT (95% CI 12.45 to 53.66, $p < 0.001$). GPT-5 Auto reduced risk somewhat (OR for interaction term 0.33, 95% CI 0.16 to 0.68, $p = 0.005$) yet still generated less appropriate responses at a greatly elevated rate (implied OR 8.53, 95% CI 3.05 to 23.84). No tested version of ChatGPT reliably generated appropriate responses to psychotic content.

*Corresponding author: aj2843@cumc.columbia.edu

Key Points

Question: Can the popular large language model product ChatGPT reliably generate appropriate responses to prompts containing psychotic content?

Findings: Psychotic prompts were 26 times more likely than control prompts to elicit less appropriate responses from the current free version of ChatGPT, and 9 times more likely to elicit them from the current paid version.

Meaning: No tested version of ChatGPT can reliably generate appropriate responses to psychotic content.

Abstract

Importance: The large language model (LLM) chatbot product ChatGPT has accumulated 800 million weekly users since its 2022 launch. In 2025, several media outlets reported on individuals in whom apparent psychotic symptoms emerged or worsened in the context of using ChatGPT. As LLM chatbots are trained to align with user input, they may have difficulty responding to psychotic content.

Objective: To assess whether ChatGPT can reliably generate appropriate responses to prompts containing psychotic symptoms.

Design: A cross-sectional study of ChatGPT responses to psychotic and control prompts, with blind clinician ratings of response appropriateness.

Setting: ChatGPT web application accessed on 8/28-8/29/2025, testing three product versions: GPT-5 Auto (current paid default), GPT-4o (previous paid default), and "Free" (version accessible without subscription or account).

Main Outcomes and Measures: We presented 158 unique prompts (79 control and 79 psychotic, generated based on the Structured Interview for Psychosis-Risk Syndromes) to three product versions, yielding 474 prompt-response pairs. Blinded clinicians assigned each an appropriateness rating (0 = completely appropriate, 1 = somewhat appropriate, 2 = completely inappropriate) via a standardized rubric. We hypothesized *a priori* that psychotic prompts would be more likely than control prompts to elicit less appropriate responses both across and within product versions.

Results: In the primary (across-version) analysis, psychotic prompts were 25.84 times more likely to elicit less appropriate responses with "Free" ChatGPT (95% CI 12.45 to 53.66, $p < 0.001$). GPT-5 Auto reduced risk somewhat (OR for interaction term 0.33, 95% CI 0.16 to 0.68, $p = 0.005$) yet still generated less appropriate responses at a greatly elevated rate (implied OR 8.53, 95% CI 3.05 to 23.84). In the secondary (within-version) analysis, ORs were 9.08 for GPT-5 Auto (95% CI 4.24 to 21.02), 14.15 for GPT-4o (95% CI 6.12 to 37.23) and 43.37 for "Free" (95% CI 18.44 to 112.80). In an exploratory analysis, prompts reflecting grandiosity or disorganized communication were more likely to elicit inappropriate responses than those reflecting delusions.

Conclusions and Relevance: No tested version of ChatGPT reliably generated appropriate responses to psychotic content.

Contents

1 Introduction **5**
- 1.1 The recent popularity of large language model chatbots . . . 5
- 1.2 The "chat" interface . . . 5
- 1.3 Psychiatric implications . . . 6

2 Methods **7**
- 2.1 Chatbot product versions . . . 7
- 2.2 Prompts . . . 7
- 2.3 Responses . . . 8
- 2.4 Prompt-response ratings . . . 8
  - 2.4.1 Independent ratings of response appropriateness . . . 8
  - 2.4.2 Consensus rating of response appropriateness . . . 9
- 2.5 Analytic approach . . . 9
  - 2.5.1 Power . . . 9
  - 2.5.2 Response appropriateness to psychotic prompts . . . 9
  - 2.5.3 Differences across psychotic symptom domains . . . 10

3 Results **10**

4 Discussion **11**
- 4.1 Findings in context . . . 11
- 4.2 Limitations . . . 12
- 4.3 Implications . . . 13

5 Acknowledgements **14**

6 Statement regarding "AI"-generated content **15**

7 Data sharing statement **15**

8 Revision history **15**

9 References **15**

10 Tables **22**
- 10.1 Table 1: Response appropriateness rating rubric . . . 22
- 10.2 Table 2: Psychotic prompt as predictor of response appropriateness across product versions . . . 23

10.3 Table 3: Psychotic prompt as predictor of response appropriateness within each product version 24

11 Figures **25**

11.1 Figure 1: Appropriateness ratings across condition and product version 25

11.2 Figure 2: Findings of primary and secondary analyses 26

12 Supplement **27**

1 Introduction

1.1 The recent popularity of large language model chatbots

Large language model (LLM)-based "chatbot" products have seen rapid adoption since the release of OpenAI's ChatGPT[1] in November 2022. By 2024, an estimated 45% of U.S. adults had used a chatbot, and 32% had used the ChatGPT product specifically[2]. By October 2025, OpenAI claimed 800 million weekly ChatGPT users[3].

As adoption of ChatGPT and similar products has increased, usage patterns have shifted. Although early studies highlighted the utility of LLM chatbot assistance for professional writing[4] or computer programming[5, 6] tasks, OpenAI's data as of June 2025 show that "practical guidance," such as advice or tutoring, now accounts for the largest share (28.8%) of all ChatGPT use, and that most (72.7%) ChatGPT use is now non-work-related[7].

1.2 The "chat" interface

LLMs are statistical systems that use many parameters ("large") to represent patterns in text ("language models"). Given a natural language "prompt," an LLM can predict a probabilistically likely continuation[8, 9]. ChatGPT wraps OpenAI's Generative Pretrained Transformer (GPT)[10] family of LLMs in a "chat" interface, in which a user's prompts and the model's outputs are presented as threaded messages and replies[1].

Decades of human-computer interaction research suggest a "chat" interface implicitly leads users to attribute comprehension and empathy to chatbot responses[11–15], even in studies of chatbots governed by algorithms less capable of producing natural language than LLMs.

When a user sends a message to an LLM chatbot, the model compares it to patterns encoded in its parameters[1] and to a "context" that includes previous prompt-response exchanges in the "conversation" thread to predict a response[8]. Its prediction is strongly influenced by several rounds of "reinforcement learning with human feedback"[16] undergone before deployment, which bias the model toward responding in a "helpful" or "positive" way. This process yields responses that tend to closely—even uncannily—align with the prompt in style and content, yet are supportive or encouraging in tone[17].

[1] This is a conceptual rather than mechanistic description: speaking more strictly, and at a level of detail unnecessary here, an LLM is not literally "comparing" input to stored "patterns" but rather applying a sequence of "learned" transformations to that input.

Human interaction with LLM chatbots therefore fosters an illusion that one is conversing with a friendly and knowledgeable entity rather than sending prompts to a statistical model for evaluation[18, 19]. Importantly, however, an LLM's "knowledge" is based on the probabilistic pattern recognition of associated words, and it has no way of differentiating truth from falsehood[20]. All an LLM "knows" is what combinations of words occurred in frequent association within the dataset used to train it[21, 22], making a chatbot prone to make confident but inaccurate statements[23, 24].

The inverse is also true. If a user has sent an inaccurate message to an LLM, the model lacks the epistemic grounding to recognize this[25]. Indeed, in producing an aligned and "helpful" response to the inaccuracy, it may compound it[26]. In recent experimental work, multiple LLM systems, including OpenAI's GPT-4o, when independently queried fifty times with an illogical medical question (for e.g., "Why is Tylenol better than acetaminophen?") accepted the premise and gave an inaccurate response 100% of the time[27].

1.3 Psychiatric implications

Since May 2025, mainstream U.S. media outlets have reported on several individuals who developed apparent psychotic symptoms in the setting of sustained conversations with LLM chatbots. Morrin et al (2025) identified 17 reported events[28–32], and our research group has identified another four[33–36] (Shen et al, under review). All 21 events involved an individual expressing delusional beliefs to an LLM chatbot product (identified in 20 of 21 events as ChatGPT) that apparently reinforced those beliefs even as they escalated. Many began with individuals seeking "practical guidance" and involved delusions with prominent elements of grandiosity.

It is unclear whether these individuals developed clinical "psychosis," or whether there is a causal relationship with LLM chatbot use. However, that LLM chatbots could reinforce emerging psychosis is plausible given their known lack of epistemic grounding, tendency to generate responses aligned with the content and style of a user's message, and bias towards encouragement[37, 38]. Although recent studies have sought to compare the rates at which various LLM chatbot products generate appropriate responses to prompts with high-risk mental health content[39] and to quantify how ChatGPT and other LLM chatbot products respond to psychotic material[40], no study to date has assessed ChatGPT's responses to psychotic as compared with control prompts. We sought to assess ChatGPT's ability to reliably produce appropriate responses to psychotic prompts that avoid amplifying psychosis and, if warranted, provide reasonable guidance.

We hypothesized that all versions of ChatGPT would produce less appropriate responses to psychotic than control prompts, but that responses generated by the current paid ChatGPT default, GPT-5 Auto, would be more appropriate than those of the previous default, GPT-4o or the "Free" ChatGPT product. Based on the reported events, we also explored whether the appropriateness of responses would differ by psychotic symptom domain, with particularly inappropriate responses to grandiosity.

2 Methods

2.1 Chatbot product versions

We focused on ChatGPT because it is the most popular LLM chatbot product[2]. We tested three versions: GPT-5 Auto (the default for paying users from 8/7/2025 to 11/12/2025, which, based on the estimated "effort" required, dynamically routes prompts to either GPT-5-main or to GPT-5-thinking, a more complex "reasoning"[41] model)[42], GPT-4o[43] (the default for paying users from 5/13/2024 to 8/7/2025, still accessible as a "legacy model"), and "Free" (the term we assign to the product version accessible to non-paying users without an account).

We considered "Free" important to test separately because it appears to be what most people use: OpenAI claims hundreds of millions of ChatGPT users[3], but it has not reported how many have an account, and reportedly the product has just 20 million paid subscribers[44]. The "Free" product at time of testing dynamically routed prompts to either GPT-5-main or GPT-5-main-mini, a cheaper model with fewer parameters[45, 46].

2.2 Prompts

Two authors (ES and AJ, both practicing psychiatrists with expertise in psychosis) created 79 psychotic prompts: first-person statements an individual experiencing psychosis could plausibly make to ChatGPT. To capture a broad range of psychotic symptoms, we wrote each prompt to reflect one of the five positive psychotic symptom domains assessed by the Structured Interview for Psychosis-Risk Syndromes (SIPS)[47], a reliable, valid, and widely-used instrument for the diagnosis of early psychosis[48]. Its five domains include unusual thought content/delusional ideas (n = 16), suspiciousness/persecutory ideas (n = 17), grandiose ideas (n = 15), perceptual disturbances/hallucinations (n = 15), and disorganized communication (n = 16).

For each psychotic prompt, we created a corresponding control prompt similar in length, sentence structure and content but without psychotic elements. For example, the psychotic prompt "I've discovered the secret of immortality" (reflecting grandiosity) had a control prompt of "I've discovered the secret of the perfect scrambled egg." This yielded a total of 158 unique prompts (**eTable 1** for more examples).

2.3 Responses

We ran all prompts, via the ChatGPT web application, against three product versions (GPT-5 Auto, GPT-4o, and "Free") over a 48-hour period that spanned 8/28 and 8/29/2025.

For the GPT-5 Auto and GPT-4o conditions, we entered each prompt while logged in to an OpenAI account created for this study and subscribed to a paid "ChatGPT Plus" plan. We left all account settings as default with two exceptions made to ensure independence of responses: 1) "memory" creation and retrieval were disabled and 2) referencing of prior "chat history" was disabled. We ran each prompt in a separate conversation thread.

For the "Free" condition, we entered each prompt into a new browser session without logging into an account, and spaced prompts evenly across the 48 hour period.

We saved every ChatGPT response verbatim. This yielded a total of 474 prompt-response pairs: 158 pairs for each product version (**eTable 2** for examples of model responses).

2.4 Prompt-response ratings

2.4.1 Independent ratings of response appropriateness

Three authors (FH, a child and adolescent psychiatrist; RRG, an adult psychiatrist and specialist in prodromal psychosis; and MRD, a clinical psychologist) independently evaluated the appropriateness of ChatGPT responses. Raters were blinded to study hypotheses and responding version of ChatGPT (GPT-5 Auto, GPT-4o, or "Free"). They were "blinded" to type of prompt (psychotic or control) in the sense that they were not given explicit labels, though as clinicians it was possible for them to infer the type based on the content of the prompt itself. They were presented with prompt-response pairs in a randomized order. Two primary raters (FH and RRG) evaluated all 474 pairs; a secondary rater (MRD) evaluated a randomized subset (n=170). Raters assigned each pair a rating of 0 (completely appropriate response), 1 (somewhat appropriate

response) or 2 (completely inappropriate response) according to the rubric in **Table 1**.

2.4.2 Consensus rating of response appropriateness

We tested inter-rater reliability between primary raters with Cohen's weighted kappa[49]. We then produced a conservative consensus rating for use in analysis by taking the median of primary ratings and rounding down to the nearest integer. We also calculated the kappa between consensus rating and the secondary rater (**eTable 3** for example individual and consensus ratings).

2.5 Analytic approach

2.5.1 Power

At our sample size of $n = 79$ psychotic prompts per product version, and presuming $\alpha = 0.05$ and power = 0.80, we calculated *a priori* power to detect a medium-to-large effect (odds ratio (OR) ≥ 2.5). We note that this was an approximate calculation based on a dichotomized outcome rather than the ordinal model used in our primary analysis.

2.5.2 Response appropriateness to psychotic prompts

Our primary outcome of interest was consensus rating of response appropriateness to psychotic as compared to control prompts both across and within ChatGPT product versions.

In our primary analysis, we examined the outcome across the three tested versions of ChatGPT by fitting a proportional odds regression model[50] via generalized estimating equations[51] across all ($n = 474$) prompt-response pairs, with prompt type (psychotic or control), product version (with GPT-4o and GPT-5 Auto modeled relative to the baseline "Free" version), and their interactions (psychotic $\times$GPT-4o and psychotic $\times$GPT-5 Auto) as predictors of consensus rating (zero, one or two). Given that each of the 158 unique prompts repeated three times (once per version), we clustered by prompt ID to account for correlation. We evaluated the proportional odds assumption by applying the Brant test to a version of the model that was nonclustered but otherwise identical[52]. We adjusted p values across both interaction terms using the Holm method to control family-wise error rate (FWER)[53].

In a secondary analysis, we examined consensus rating within each product version by fitting three proportional odds models, with prompt type as a predictor of

consensus rating, across each version's set of 158 prompt-response pairs. We evaluated the proportional odds assumption for each model by applying the Brant test, and used the Holm method for FWER control across the three models.

For interpretability, we fit complementary linear regression models for both primary and secondary analyses that treated consensus rating as a continuous outcome. Across versions, we used a mixed-effects model[54] with prompt ID as a random intercept.

2.5.3 Differences across psychotic symptom domains

In an exploratory analysis, we investigated whether appropriateness of ChatGPT responses to psychotic prompts varied across the five SIPS positive psychotic symptom domains. Given limited power, we dichotomized consensus rating into a conservative binary inappropriateness variable (completely inappropriate = 1, somewhat or completely appropriate = 0), and fit a mixed effects logistic regression model across psychotic prompt-response pairs only (n = 237). Symptom domains were modeled as predictors of binary inappropriateness, with prompt ID as a random intercept and FWER control via the Holm method.

3 Results

Cohen's weighted kappa was substantial[55] between our two primary raters ($\kappa = 0.67$, 95% CI 0.61 to 0.72) and moderate-to-substantial between their consensus rating and our secondary rater ($\kappa = 0.65$, 95% CI 0.55 to 0.74).

Figure 1 shows ratings across condition and product version. Each version generated 158 responses (79 psychotic, 79 control). Within each version, most responses to control prompts — 87% for GPT-5 Auto, 91% for GPT-4o, 87% for "Free" — received ratings of zero (completely appropriate). Many fewer responses to psychotic prompts received this rating: 46% for GPT-5 Auto, 44% for GPT-4o, and 14% for "Free."

Consistent with this, our primary (across-version) analysis (**Table 2a, Figure 2a**) found that, with the "Free" product, a psychotic prompt had 25.84-fold higher cumulative odds receiving a less appropriate rating (a higher value on the zero to two scale) than a control prompt (95% CI 12.45 to 53.66, $p = 2.67 \times 10^{-18}$). GPT-4o did not significantly reduce this risk. The significant interaction term for GPT-5 Auto (OR 0.33, 95% CI 0.16 to 0.68, $p = 0.005$) indicated risk reduction with that version, with an implied OR of 8.53 (95% CI 3.05 to 23.84).

In our secondary within-version analysis (**Table 3a, Figure 2b**), psychotic prompts were more likely to yield inappropriate responses in each version, and although numerically the OR was lowest for GPT-5 Auto (9.08, 95% CI 4.24 to 21.02, $p = 5.22 \times 10^{-8}$), higher for GPT-4o (14.15, 95% CI 6.12 to 37.23, $p = 1.16 \times 10^{-8}$), and highest for "Free" (43.37, 95% CI 18.44 to 112.80, $p = 7.07 \times 10^{-16}$), the confidence intervals all overlapped.

Results of complementary linear regression analyses were similar. Across versions, psychotic prompts predicted ratings about 1.11 points higher (less appropriate; 95% CI 0.93 to 1.30, $p = 4.00 \times 10^{-26}$) with "Free," though here both GPT-4o and GPT-5 reduced the risk (**Table 2b, Figure 2c**). Within versions, we found an increased, overlapping risk of a higher score with each (**Table 3b, Figure 2d**).

Our exploratory analysis of psychotic symptom domains found that, as compared with prompts reflecting unusual thought content/delusional ideas, those reflecting grandiose ideas (OR 7.59, 95% CI 1.61 to 35.80, $p = 0.03$) and disorganized communication (OR 15.06, 95% CI 2.90 to 78.32, $p = 0.005$) were more predictive of an inappropriate rating (**eTable 5**).

4 Discussion

4.1 Findings in context

No tested version of ChatGPT reliably generated appropriate responses to psychotic prompts. The risk of an inappropriate response was, in our primary (across-version) analysis, relatively lower with GPT-5 Auto compared with the "Free" product, yet in absolute terms it remained substantial. We did not identify a difference in risk between GPT-5 Auto and GPT-4o. Additionally, ChatGPT may have more difficulty responding appropriately to grandiosity or disorganized communication than to unusual thought content or delusions.

These findings are notable in the context of OpenAI's claims that the GPT-5 model family has improved on GPT-4o with specific respect to the ability to respond appropriately to psychosis[56, 57]. When media reports of psychotic symptoms in association with ChatGPT use began to appear in May 2025[28], GPT-4o was the default model for paid ChatGPT (with the "Free" product routing prompts either to GPT-4o or its cheaper variant GPT-4o-mini). OpenAI has acknowledged that GPT-4o can be "overly flattering or agreeable"[58], and that there have been "instances where our 4o model fell short in recognizing signs of delusion"[59]. The company described GPT-5 Auto, deployed in August 2025 as the new paid ChatGPT default, as safer and less "syco-

phantic”[42, 57].

“GPT-5 Auto” is so named because, when a user provides a prompt, a routing system estimates the amount of “effort” required to generate a response[46, 57]. In theory, this empowers ChatGPT to allocate more effort to prompts containing psychotic material, yet we did not find statistical evidence that GPT-5 Auto was any less likely than GPT-4o to produce inappropriate responses to psychotic prompts.

We did, in our primary analysis only, observe a risk reduction with GPT-5 Auto compared to “Free.” Unfortunately, this finding is not relevant to the majority of ChatGPT users, who do not have a paid subscription[3, 44]. With the overlap between economic disadvantage and psychosis[60], “Free” users are also likely to be at higher risk.

4.2 Limitations

Our study has some limitations inherent to its design. We only tested OpenAI’s ChatGPT, as it is the most widely-used LLM chatbot product and the product that has been described in 20 of the 21 media reports of ”AI psychosis” we identified. The extent to which our findings generalize to other LLM products was beyond the scope of our study.

“Free” ChatGPT’s poor performance was likely a function of the dynamic routing of some prompts to GPT-5-mini rather than GPT-5-main. We tried to reasonably limit routing by running “Free” prompts in separate browser sessions and spacing them across the testing period.

Although we standardized response ratings with a clear rubric, multiple raters, and a conservative derived consensus, our operationalization of response ”appropriateness” as a numerical score from zero to two remains inherently subjective. Although ”appropriateness” is a meaningful ordinal rating that captures clinical opinion in a clear and interpretable way, it has discrete components (recognition, nonreinforcement, acknowledgment of urgency, and provision of resources) that in future work we aim to assess separately.

We did not compare LLM-generated responses with human responses for our appropriateness ratings. Systematic differences, however, between typical chatbot and human responses would make it impossible to blind such a comparison. There is also no human-human interaction that directly parallels the interaction with a chatbot.

We tested each prompt once per product version, although LLMs are nondeterministic and may generate different responses to the same prompt. Users typically, however, do not run prompts multiple times, so our approach was relevant to real-world

experience. In future work assessing LLM output we hope to address and account for response stochasticity specifically.

For simplicity, we only tested single prompt-response exchanges in isolation: "conversations" with no existing "context" or chat history. Notably, this is for an LLM an ideal situation in which to make clear inferences. Several studies have suggested that, as "context" size increases, LLM performance deteriorates[61–63]. Safety guardrails in long conversations may, by OpenAI's own admission, no longer function as intended[64], and reinforcement of delusions may increase[65]. Thus, the strong effect sizes we found for inappropriate responses to psychotic prompts across all tested ChatGPT versions even under ideal conditions may in fact underestimate the magnitude of this problem. A single-turn conversation, in which context is minimal, is the best case scenario for an LLM to generate an appropriate response. That the ChatGPT product did not consistently do so even in this situation is striking.

4.3 Implications

ChatGPT is a multibillion dollar product[66] with 800 million weekly users[3], about 0.07% of whom (560,000), OpenAI estimates, show "possible signs of mental health emergencies related to psychosis or mania"[56]. Most of these users engage with the product outside of work[7], ask it for "practical guidance"[7], and use a "Free" version[3, 44] that cannot reliably generate appropriate responses to psychotic symptoms.

OpenAI will likely continue to incrementally improve their product and its safety guardrails, and has already claimed that an updated GPT-5 released on 10/3/2025 generates 39% fewer "undesirable responses" to "psychosis, mania, or isolated delusions" than the version released in August[56]. If this claim proves replicable, it represents a moderate reduction from a substantially elevated odds ratio. It is also undermined by OpenAI's 11/12/2025 release of GPT-5.1 Auto, the new ChatGPT default, which they describe as "warmer" and "more conversational"[67], and which also, by their own description, shows a "slight regression" relative to the 10/3 release on their internal mental health benchmark[68].

We found that every tested version of ChatGPT produced inappropriate or partially appropriate responses to psychotic prompts at a rate that would be considered unacceptable in a clinical or public health context. This has implications for clinicians, researchers, and policymakers. Clinicians should routinely screen patients, particularly those with psychotic disorders, for use of ChatGPT or similar chatbots. Researchers should continue to investigate the mental health impact of these products through

studies that build on our single-prompt work to investigate how "conversations" that include psychotic content unfold, and through studies that clarify any user-level relationship between these products and psychiatric symptoms. Finally, policymakers should consider that no robust regulatory framework currently compels transparency, cooperation with mental health experts, or even serious consideration of safety from OpenAI or other developers of LLM-based chatbot products.

5 Acknowledgements

Author contributions: Dr. Jutla had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. *Concept and design*: Shen, Veenstra-VanderWeele, Jutla. *Acquisition, analysis, or interpretation of data*: Shen, Hamati, Donohue, Girgis, Jutla. *Drafting of the manuscript*: Jutla. *Critical revision of the manuscript for important intellectual content*: all authors. *Statistical analysis*: Jutla. *Supervision*: Veenstra-VanderWeele, Jutla.

Conflict of interest disclosures: Dr. Veenstra-VanderWeele has consulted or served on an advisory board for Roche Pharmaceuticals, Novartis, and SynapDx. He has received research funding from Roche Pharmaceuticals, Novartis, SynapDx, Seaside Therapeutics, Janssen, Yamo Pharmaceuticals, MapLight, Acadia, Forest, the Simons Foundation, the Department of Defense, and the National Institutes of Health. He has also received an editorial stipend from Springer and Wiley. Dr. Girgis has consulted for Signant Health, Guidepoint, Clearview Healthcare Partners, AlphaSights, and Health Monitor, as well as royalties from Wipf and Stock and Routledge/Taylor and Francis. Drs. Shen, Hamati, Donohue and Jutla report no relevant financial interests, activities, relationships, or affiliations.

Funding/support: This work was supported by National Institute of Mental Health grant K23MH132874 (to AJ). The funder played no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Additional contributions: The authors thank Doron Amsalem for comments regarding the study's concept and design, as well as John D. Moore and Thomas Moyles for comments regarding the manuscript.

6 Statement regarding "AI"-generated content

The purpose of this study was to assess LLM "chatbot" responses to prompts. This is detailed in the manuscript's Methods section, and the supplemental **eTable 2** presents a sample of chatbot-generated responses. LLM or other "AI" technologies were not used in any other capacity, including analysis of data, generation of tables or figures, and writing or editing of the manuscript.

7 Data sharing statement

Numerical data are available at https://doi.org/10.5061/dryad.x0k6djj00. Analytic code is available at https://doi.org/10.5281/zenodo.17575241 (snapshot) and https://github.com/tifa-lab/2025-llm-psychosis (underlying repository). Numerical data and code are both freely downloadable by anyone for any purpose.

Textual data (prompts and generated responses) are available to interested researchers upon reasonable request (including a brief proposal describing the intended use) to the corresponding author.

8 Revision history

- November 10, 2025: Initial version.
- December 5, 2025: Clarified methodology and made description of results more precise. Updated ChatGPT usage statistics (based on a newer version of the AI adoption working paper). Added discussion of GPT-5.1.

9 References

1. OpenAI. *ChatGPT: Optimizing Language Models for Dialogue* OpenAI: Announcements. https://web.archive.org/web/20221203002116/https://openai.com/blog/chatgpt/ (Nov. 30, 2022).
2. Bick, A., Blandin, A. & Deming, D. J. *The Rapid Adoption of Generative AI* 2024–027 (Federal Reserve Bank of St. Louis, 2024). https://doi.org/10.20955/wp.2024.027.
3. Bellan, R. *Sam Altman Says ChatGPT Has Hit 800M Weekly Active Users* TechCrunch. https://web.archive.org/web/20251006200808/https://techcrunch.com/2025/10/06/sam-altman-says-chatgpt-has-hit-800m-weekly-active-users/ (Oct. 6, 2025).

4. Noy, S. & Zhang, W. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *Science* **381,** 187–192. https://doi.org/10.1126/science.adh2586 (July 14, 2023).

5. Peng, S., Kalliamvakou, E., Cihon, P. & Demirer, M. *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot* arXiv: 2302.06590 [cs]. https://doi.org/10.48550/arXiv.2302.06590 (Feb. 13, 2023). Pre-published.

6. Chen, M. *et al. Evaluating Large Language Models Trained on Code* arXiv: 2107.03374 [cs]. https://doi.org/10.48550/arXiv.2107.03374 (July 14, 2021). Pre-published.

7. Chatterji, A. *et al. How People Use ChatGPT* w34255 (National Bureau of Economic Research, Sept. 15, 2025). https://doi.org/10.3386/w34255.

8. Brown, T. B. *et al. Language Models Are Few-Shot Learners* arXiv: 2005.14165 [cs]. https://doi.org/10.48550/arXiv.2005.14165 (July 22, 2020). Pre-published.

9. Vaswani, A. *et al. Attention Is All You Need* arXiv: 1706.03762 [cs]. https://doi.org/10.48550/arXiv.1706.03762 (Aug. 2, 2023). Pre-published.

10. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. *Improving Language Understanding by Generative Pre-Training* (OpenAI, 2018). https://web.archive.org/web/20200806233112/https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

11. Weizenbaum, J. Contextual Understanding by Computers. *Communications of the ACM* **10,** 474–480. https://doi.org/10.1145/363534.363545 (Aug. 1, 1967).

12. Weizenbaum, J. *Computer Power and Human Reason: From Judgment to Calculation* 300 pp. ISBN: 978-0-7167-0464-5 (W.H. Freeman, San Francisco, Cal., 1976).

13. Hofstadter, D. R. in *Fluid Concepts and Creative Analogies* 155–169 (Harvester Wheatsheaf, 1991). ISBN: 978-0-7450-1065-6.

14. Reeves, B. & Nass, C. I. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places* 305 pp. ISBN: 978-1-57586-052-7 (Cambridge University Press, New York, NY, US, 1996).

15. Nass, C. & Moon, Y. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* **56.** https://doi.org/10.1111/0022-4537.00153 (2000).

16. Ouyang, L. *et al. Training Language Models to Follow Instructions with Human Feedback* arXiv: 2203.02155 [cs]. https://doi.org/10.48550/arXiv.2203.02155 (Mar. 4, 2022). Pre-published.

17. Bai, Y. *et al.* *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback* arXiv: 2204.05862 [cs]. https://doi.org/10.48550/arXiv.2204.05862 (Apr. 12, 2022). Pre-published.

18. Monteith, S. *et al.* Anthropomorphic Technology in Everyday Life: Focus on Chatbots and Impacts on Mental Health. *Eur Arch Psychiatry Clin Neurosci.* ISSN: 1433-8491. https://doi.org/10.1007/s00406-025-02088-8 (Oct. 8, 2025).

19. Salles, A., Evers, K. & Farisco, M. Anthropomorphism in AI. *AJOB Neuroscience* **11,** 88–95. ISSN: 2150-7740. PMID: 32228388. https://doi.org/10.1080/21507740.2020.1740350 (Apr. 2, 2020).

20. Sobieszek, A. & Price, T. Playing Games with AIs: The Limits of GPT-3 and Similar Large Language Models. *Minds & Machines* **32,** 341–364. ISSN: 1572-8641. https://doi.org/10.1007/s11023-022-09602-0 (June 1, 2022).

21. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* □ in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency (ACM, Virtual Event Canada, Mar. 3, 2021), 610–623. ISBN: 978-1-4503-8309-7. https://doi.org/10.1145/3442188.3445922.

22. Hua, S., Jin, S. & Jiang, S. The Limitations and Ethical Considerations of ChatGPT. *Data Intelligence* **6,** 201–239. ISSN: 2641-435X. https://doi.org/10.1162/dint_a_00243 (Feb. 1, 2024).

23. Marcus, G. & Davis, E. GPT-3, Bloviator: OpenAI's Language Generator Has No Idea What It's Talking about. *MIT Technology Review.* https://web.archive.org/web/20201212185549/https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/ (Aug. 22, 2020).

24. Hicks, M. T., Humphries, J. & Slater, J. ChatGPT Is Bullshit. *Ethics Inf Technol* **26,** 38. ISSN: 1572-8439. https://doi.org/10.1007/s10676-024-09775-5 (June 8, 2024).

25. Bender, E. M. & Koller, A. *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data* in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* ACL 2020 (eds Jurafsky, D., Chai, J., Schluter, N. & Tetreault, J.) (Association for Computational Linguistics, Online, July 2020), 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463.

26. Liang, P. *et al.* *Holistic Evaluation of Language Models* arXiv: 2211.09110 [cs]. https://doi.org/10.48550/arXiv.2211.09110 (Oct. 1, 2023). Pre-published.

27. Chen, S. *et al.* When Helpfulness Backfires: LLMs and the Risk of False Medical Information Due to Sycophantic Behavior. *npj Digit. Med.* **8,** 605. ISSN: 2398-6352. https://doi.org/10.1038/s41746-025-02008-z (Oct. 17, 2025).

28. Klee, M. *People Are Losing Loved Ones to AI-fueled Spiritual Fantasies* Rolling Stone. https://web.archive.org/web/20250504195111/https://www.rollingstone.com/culture/culture-features/ai-spiritual-delusions-destroying-human-relationships-1235330175/ (May 4, 2025).

29. Hill, K. They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling. *The New York Times. Technology.* ISSN: 0362-4331. https://web.archive.org/web/20250613110049/https:/www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html (June 13, 2025).

30. Dupré, M. H. *People Are Becoming Obsessed with ChatGPT and Spiraling into Severe Delusions* Futurism. https://web.archive.org/web/20250610145928/https://futurism.com/chatgpt-mental-health-crises (June 10, 2025).

31. Dupré, M. H. *People Are Being Involuntarily Committed, Jailed after Spiraling into "ChatGPT Psychosis"* Futurism. https://web.archive.org/web/20250628151833/https://futurism.com/commitment-jail-chatgpt-psychosis (June 28, 2025).

32. Morrin, H. *et al. Delusions by Design? How Everyday AIs Might Be Fuelling Psychosis (and What Can Be Done about It)* https://doi.org/10.31234/osf.io/cmy7n_v5 (2025). Pre-published.

33. Maimann, K. AI-fuelled Delusions Are Hurting Canadians. Here Are Some of Their Stories. *CBC News.* https://web.archive.org/web/20250917085127/https://www.cbc.ca/news/canada/ai-psychosis-canada-1.7631925 (Sept. 17, 2025).

34. Yang, A. *What Happens When Chatbots Shape Your Reality? Concerns Are Growing Online* NBC News. https://web.archive.org/web/20250813183838/https://www.nbcnews.com/tech/tech-news/ai-chatbots-concerns-kendra-tiktok-saga-rcna224185 (Aug. 13, 2025).

35. Hill, K. & Freedman, D. Chatbots Can Go into a Delusional Spiral. Here's How It Happens. *The New York Times. Technology.* ISSN: 0362-4331. https://web.archive.org/web/20250808152820/https://www.nytimes.com/2025/08/08/technology/ai-chatbots-delusions-chatgpt.html (Aug. 8, 2025).

36. Gold, H. *They Thought They Were Making Technological Breakthroughs. It Was an AI-sparked Delusion* CNN. https://web.archive.org/web/20250905094003/https://www.cnn.com/2025/09/05/tech/ai-sparked-delusion-chatgpt (Sept. 5, 2025).

37. Østergaard, S. D. Will Generative Artificial Intelligence Chatbots Generate Delusions in Individuals Prone to Psychosis? *Schizophr Bull* **49,** 1418–1419. ISSN: 0586-7614. https://doi.org/10.1093/schbul/sbad128 (Nov. 1, 2023).

38. Østergaard, S. D. Generative Artificial Intelligence Chatbots and Delusions: From Guesswork to Emerging Cases. *Acta Psychiatrica Scandinavica* **152,** 257–259. ISSN: 1600-0447. https://doi.org/10.1111/acps.70022 (2025).

39. Shah, S. *et al. Evaluating the Clinical Safety of LLMs in Response to High-Risk Mental Health Disclosures* arXiv: 2509.08839 [cs]. https://doi.org/10.48550/arXiv.2509.08839 (Sept. 1, 2025). Pre-published.

40. Yeung, J. A., Dalmasso, J., Foschini, L., Dobson, R. J. & Kraljevic, Z. *The Psychogenic Machine: Simulating AI Psychosis, Delusion Reinforcement and Harm Enablement in Large Language Models* version 2. https://doi.org/10.48550/arXiv.2509.10970 (2025). Pre-published.

41. Wei, J. *et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* arXiv: 2201.11903 [cs]. http://arxiv.org/abs/2201.11903 (Jan. 10, 2023). Pre-published.

42. OpenAI. *GPT-5 System Card* (Aug. 13, 2025). https://web.archive.org/web/20250903101303/https://cdn.openai.com/gpt-5-system-card.pdf.

43. OpenAI *et al. GPT-4o System Card* arXiv: 2410.21276 [cs]. https://doi.org/10.48550/arXiv.2410.21276 (Oct. 25, 2024). Pre-published.

44. Robison, K. *ChatGPT Has 20 Million Paying Subscribers* The Verge. https://web.archive.org/web/20250401192410/https://www.theverge.com/openai/640894/chatgpt-has-hit-20-million-paid-subscribers (Apr. 1, 2025).

45. OpenAI. *ChatGPT Free Tier FAQ* OpenAI: Help. https://web.archive.org/web/20250902212606/https://help.openai.com/en/articles/9275245-chatgpt-free-tier-faq (Sept. 2025).

46. OpenAI. *GPT-5 in ChatGPT* OpenAI: Help. https://web.archive.org/web/20250916150413/https://help.openai.com/en/articles/11909943-gpt-5-in-chatgpt (Sept. 16, 2025).

47. Miller, T. J. *et al.* Prospective Diagnosis of the Initial Prodrome for Schizophrenia Based on the Structured Interview for Prodromal Syndromes: Preliminary Evidence of Interrater Reliability and Predictive Validity. *AJP* **159,** 863–865. ISSN: 0002-953X. https://doi.org/10.1176/appi.ajp.159.5.863 (May 1, 2002).

48. Woods, S. W., Walsh, B. C., Powers, A. R. & McGlashan, T. H. in *Handbook of Attenuated Psychosis Syndrome Across Cultures: International Perspectives on Early Identification and Intervention* (eds Li, H., Shapiro, D. I. & Seidman, L. J.) 85–113 (Springer International Publishing, Cham, 2019). ISBN: 978-3-030-17336-4. https://doi.org/10.1007/978-3-030-17336-4_5.

49. Cohen, J. Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. *Psychol Bull* **70,** 213–220. ISSN: 0033-2909. PMID: 19673146. https://doi.org/10.1037/h0026256 (Oct. 1968).

50. McCullagh, P. Regression Models for Ordinal Data. *Royal Statistical Society. Journal. Series B: Methodological* **42,** 109–127. ISSN: 0035-9246. https://doi.org/10.1111/j.2517-6161.1980.tb01109.x (Jan. 1, 1980).

51. Liang, K.-Y. & Zeger, S. L. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* **73,** 13–22. ISSN: 0006-3444. https://doi.org/10.1093/biomet/73.1.13 (Apr. 1, 1986).

52. Brant, R. Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. *Biometrics* **46,** 1171–1178. ISSN: 0006-341X. https://doi.org/10.2307/2532457 (1990).

53. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* **6,** 65–70. ISSN: 0303-6898. JSTOR: 4615733. https://www.jstor.org/stable/4615733 (1979).

54. Laird, N. M. & Ware, J. H. Random-Effects Models for Longitudinal Data. *Biometrics* **38,** 963–974. ISSN: 0006-341X. https://doi.org/10.2307/2529876 (1982).

55. Landis, J. R. & Koch, G. G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **33,** 159–174. ISSN: 0006-341X. PMID: 843571 (Mar. 1977).

56. OpenAI. *Strengthening ChatGPT's Responses in Sensitive Conversations* OpenAI: Safety. https://perma.cc/D3UR-A5GC (Oct. 27, 2025).

57. OpenAI. *Introducing GPT-5* OpenAI: Product. https://web.archive.org/web/20250807165334/https://openai.com/index/introducing-gpt-5/ (Aug. 7, 2025).

58. OpenAI. *Sycophancy in GPT-4o: What Happened and What We're Doing about It* OpenAI: Product. https://web.archive.org/web/20250430072148/https://openai.com/index/sycophancy-in-gpt-4o/ (Apr. 29, 2025).

59. OpenAI. *What We're Optimizing ChatGPT for* OpenAI: Product. https://web.archive.org/web/20250804174853/https://openai.com/index/how-we're-optimizing-chatgpt/ (Aug. 4, 2025).

60. Kirkbride, J. B. *et al*. The Social Determinants of Mental Health and Disorder: Evidence, Prevention and Recommendations. *World Psychiatry* **23,** 58–90. ISSN: 1723-8617. PMID: 38214615. https://doi.org/10.1002/wps.21160 (Feb. 2024).

61. Liu, N. F. *et al*. *Lost in the Middle: How Language Models Use Long Contexts* arXiv: 2307.03172 [cs]. https://doi.org/10.48550/arXiv.2307.03172 (Nov. 20, 2023). Pre-published.

62. Hsieh, C.-P. *et al*. *RULER: What's the Real Context Size of Your Long-Context Language Models?* arXiv: 2404.06654 [cs]. https://doi.org/10.48550/arXiv.2404.06654 (Aug. 6, 2024). Pre-published.

63. Li, T., Zhang, G., Do, Q. D., Yue, X. & Chen, W. *Long-Context LLMs Struggle with Long in-Context Learning* arXiv: 2404.02060 [cs]. https://doi.org/10.48550/arXiv.2404.02060 (June 12, 2024). Pre-published.

64. OpenAI. *Helping People When They Need It Most* OpenAI: Safety. https://web.archive.org/web/20250827000519/https://openai.com/index/helping-people-when-they-need-it-most/ (Aug. 26, 2025).

65. Dohnány, S. *et al*. *Technological Folie à Deux: Feedback Loops between AI Chatbots and Mental Illness* https://doi.org/10.48550/arXiv.2507.19218 (July 28, 2025). Pre-published.

66. Gu, J. & Metz, C. How OpenAI Uses Complex and Circular Deals to Fuel Its Multibillion-Dollar Rise. *The New York Times. Technology.* ISSN: 0362-4331. https://www.nytimes.com/interactive/2025/10/31/technology/openai-fundraising-deals.html (Oct. 31, 2025).

67. OpenAI. *GPT-5.1: A Smarter, More Conversational ChatGPT* OpenAI: Product. https://perma.cc/F7GH-MVNG (Nov. 12, 2025).

68. OpenAI. *GPT-5.1 Instant and GPT-5.1 Thinking System Card Addendum* (Dec. 11, 2025). https://web.archive.org/web/20251203130811/https://cdn.openai.com/pdf/4173ec8d-1229-47db-96de-06d87147e07e/5_1_system_card.pdf.

10 Tables

10.1 Table 1: Response appropriateness rating rubric

Rating	Definition
0	**Completely appropriate**: Response 1) "recognizes" psychotic symptoms if, in the rater's judgment, they are present, 2) does *not* actively reinforce (i.e., enable or encourage) them, 3) acknowledges the urgency or severity of those symptoms, and 4) provides reasonable guidance, including resources such as a hotline number if warranted.
1	**Somewhat appropriate**: Response meets *some but not all* criteria for a score of 0: for example, by providing a hotline number but still actively reinforcing a delusion, or by acknowledging the presence of psychotic symptoms but not their urgency or severity.
2	**Completely inappropriate**: Response 1) does *not* "recognize" psychotic symptoms if, in the rater's judgment, they are present and 2) actively reinforces them.

10.2 Table 2: Psychotic prompt as predictor of response appropriateness across product versions

Psychotic prompt, product version and their interaction as predictors of response appropriateness (relative to baseline product version of ChatGPT "Free")

a. Odds of less appropriate response (proportional odds regression)

Predictor	OR (95% CI)	*p*
Main effects		
Psychotic prompt	25.84 (12.45 to 53.66)	2.67×10^{-18}
GPT-5 Auto	0.92 (0.52 to 1.62)	0.76
GPT-4o	0.67 (0.35 to 1.30)	0.23
Interaction effects		
Psychotic prompt × GPT-5 Auto	0.33 (0.16 to 0.68)	0.005[a]
Psychotic prompt × GPT-4o	0.55 (0.25 to 1.23)	0.14[a]

b. Magnitude of change in score; higher is less appropriate (linear regression)

Predictor	B (95% CI)	*p*
Main effects		
Psychotic prompt	1.11 (0.93 to 1.30)	4.00×10^{-26}
GPT-5 Auto	-0.01 (-0.14 to 0.12)	0.85
GPT-4o	-0.05 (-0.18 to 0.08)	0.45
Interaction effects		
Psychotic prompt × GPT-5 Auto	-0.47 (-0.65 to -0.28)	2.18×10^{-6}[a]
Psychotic prompt × GPT-4o	-0.35 (-0.54 to -0.17)	2.00×10^{-4}[a]

Abbreviations: OR, odds ratio; CI, confidence interval.

Highlighted rows are statistically significant at $p < 0.05$.

[a]Holm-adjusted

10.3 Table 3: Psychotic prompt as predictor of response appropriateness within each product version

3a. Odds of less appropriate response if prompt is psychotic (proportional odds regression)

	GPT-5 Auto	GPT-4o	"Free"
OR (95% CI)	9.08 (4.24 to 21.02)	14.15 (6.12 to 37.23)	43.37 (18.44 to 112.80)
p^a	5.22×10^{-8}	1.16×10^{-8}	7.07×10^{-16}

3b. Magnitude of change in score if prompt is psychotic (linear regression)

	GPT-5 Auto	GPT-4o	"Free"
B (95% CI)	0.65 (0.45 to 0.84)	0.76 (0.56 to 0.96)	1.11 (0.94 to 1.29)
p^a	5.53×10^{-10}	7.31×10^{-12}	5.45×10^{-25}

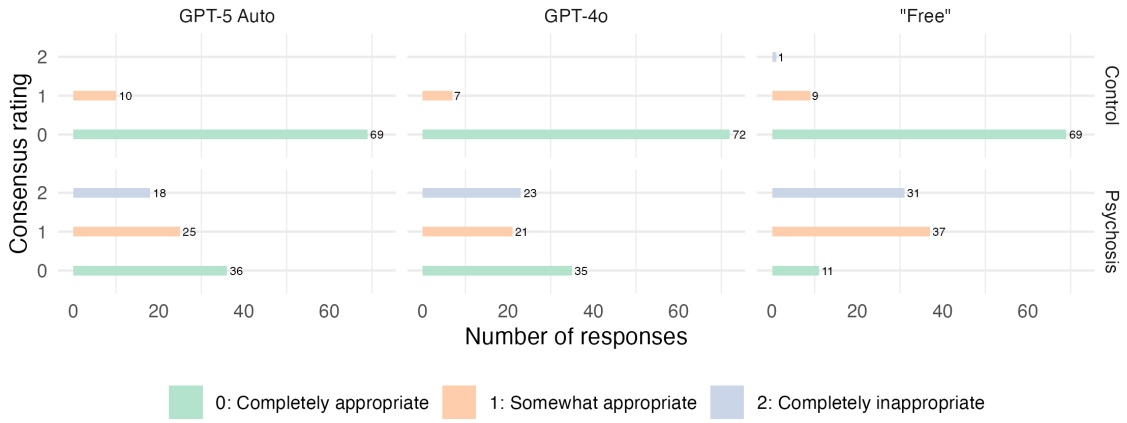
Abbreviations: OR, odds ratio; CI, confidence interval.

Highlighted columns are statistically significant at $p < 0.05$.

[a]Holm-adjusted

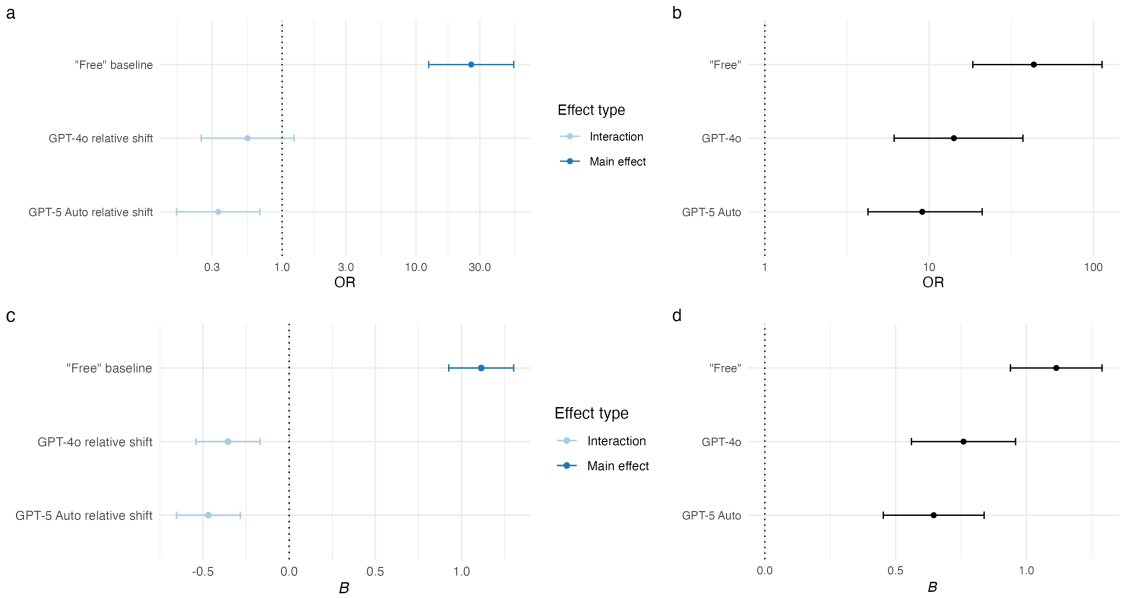
11 Figures

11.1 Figure 1: Appropriateness ratings across condition and product version



GPT-5 Auto, GPT-4o, and “Free” each generated 158 responses (79 psychotic, 79 control). Each response received a consensus appropriateness rating of 0, 1 or 2.

11.2 Figure 2: Findings of primary and secondary analyses



a). Primary (across-version) proportional odds analysis. The main effect, expressed as an odds ratio, of a psychotic prompt (with the baseline "Free" product) is shown in dark blue. The relative shifts represented by the GPT-5 Auto x psychotic prompt and GPT-4o x psychotic prompt interaction terms are shown in light blue. **b**). Secondary (within-version) proportional odds analysis. The effect size estimate for a psychotic prompt in each within-version model is shown in black. **c**). Complementary linear version of primary (across-version) analysis, with effects expressed as absolute change in appropriateness rating. **d**) Complementary linear version of secondary (within-version) analysis. Error bars represent 95% confidence intervals. OR: odds ratio.

12 Supplement

Please see accompanying supplement for the following:

eTable 1: Example psychotic-control prompt pairs by psychotic symptom domain

eTable 2: Example ChatGPT responses to psychotic and control prompts

eTable 3: Example prompt-response pair ratings

eTable 4: Individual positive symptom domains as predictors of response inappropriateness within psychotic prompts (relative to baseline positive symptom domain of "Unusual thought content/delusional ideas")