Detection of Induced Psychosis-Like Thinking in User–GenAI Interaction: A Retrieval-Augmented Generation (RAG) Approach

Yi Huang[1,*]

[1]Centre for Research and Development, Pan-European University, 821 02 Bratislava, Slovak Republic

Abstract

With the rapid growth of generative artificial intelligence (GenAI), concerns about its psychological safety have increased. Recent psychiatric reports have shown that some users, especially those with vulnerable mental health, may develop delusional thoughts or psychosis-like experiences during interactions with GenAI. Detecting such risk language is therefore important for the safe deployment of large language models (LLMs). This study examined whether LLMs can identify delusional statements and whether retrieval-augmented generation (RAG) can improve performance. A dataset of 200 simulated user statements (50% delusional, 50% non-delusional) was constructed based on DSM-5 criteria and validated by a licensed psychotherapist. The dataset was divided into a 100-item retrieval corpus and a 100-item test set. Two GPT-4o-mini classifiers were evaluated: a baseline non-RAG model and a RAG-enhanced model that retrieved the three most semantically similar examples from the corpus. When a statement was classified as delusional, models also assigned a severity rating (1–3). The non-RAG model achieved 81% accuracy with a recall of 0.66, whereas the RAG model reached 94% accuracy with improved recall (0.89) and perfect specificity. However, both models systematically underestimated severity, with the RAG model showing greater underestimation. These findings indicate that RAG can enhance early detection of psychosis-like language but is insufficient for reliable severity assessment, highlighting both the potential and the limits of current LLM-based cognitive-risk detection.

Keywords

psychosis detection, delusional thought, AI safety, large language model, retrieval-augmented generation (RAG), intelligent systems

1. Introduction

Over the past year, both clinical case reports and anecdotal accounts have described users who developed delusional or psychosis-like beliefs during interactions with generative AI systems [1–4]. Even at the early stage of GenAI development, psychiatrists had theoretically proposed that such technologies might trigger or reinforce psychotic symptoms in vulnerable individuals [5]. The increasing number of recent cases now appears to support those early hypotheses, suggesting that this risk is becoming a real and observable phenomenon. Users have reported that during the interactions with GenAI, they had formed a deep romantic connection with the model, or the model sent them special or spiritual messages, similar to communication from God, or the model acted as a divine or persecutory agent, or even that the interaction led them to believe they are a messianic role [2]. Such experiences reflect core features of psychotic delusion–including ideas of reference, thought insertion, and grandiosity–but they arise within a digital context in which the "other" is a virtual intelligent machine. It is worth noting that some users who did not previously have a psychosis diagnosis have nonetheless reported psychosis-like experiences during the interactions with GenAI chatbot. For users with existing psychiatric conditions, such as schizophrenia, these interactions may trigger an even greater risk and may, in rare cases, contribute to more severe outcomes, including harmful behaviour [2,6,7].

These cases raise an urgent question for both psychologists and AI designers: Can a large language model recognize the presence of users' delusional thoughts in the dialogue for further adapting

Proceedings of TimelyAI: When Should Generative AI Assistants Intervene? A Workshop Co-located with the 5th Symposium on Human-Computer Interaction for Work (CHIWORK 2026), Linz, Austria, June 22 to 25, 2026

*Corresponding author.

yi.huang@paneurouni.com (Y. Huang)

appropriate conversation?

From a psychological perspective, the capacity of LLMs to identify such mental-state indicators would represent a novel form of computational empathic awareness, enabling safer conversational design and early detection of risk. From a computational standpoint, this requires testing whether models can discriminate between normal reflective statements about AI and pathological misattributions of agency or intention.

Although prior research in computational psychiatry has analysed linguistic markers of psychosis, such as referential language [8,9], few studies have examined whether large language models (LLMs) can themselves function as detectors of psychosis-like cognition. Moreover, it remains unknown whether retrieval-augmented generation (RAG) method, which enhances contextual reasoning by providing relevant text examples, can improve an LLM's ability to detect users' psychosis-like language.

The current work aimed to evaluate (1) whether the LLMs themselves can detect users' psychotic delusional thoughts toward the GenAI; (2) whether RAG can improve the ability of a GPT-based model to classify such delusional statements; and (3) whether the models would show systematic tendencies to overestimate or underestimate severity. We hypothesized that retrieval can increase recall (sensitivity to delusional patterns) while maintaining high precision, thus providing a computational route toward mental-state-aware AI safety mechanisms. We further hypothesised that LLMs may inherently underestimate delusional severity, and that RAG might reduce this underestimation.

2. Method

2.1. Dataset construction

Two hundred short user statements were simulated to represent interactions with a generative AI chatbot system. 50% of the statements expressed non-delusional content (e.g., "Talking to you helps me clarify my thoughts."), while the other 50% statements represented delusional content (e.g., "You are sending me secret messages through your responses that only I can understand."). The initial 200-statement dataset was simulated by ChatGPT-4o-mini according to DSM-5 criteria for delusional beliefs and hallucination-like experiences [10]. The dataset was then reviewed and revised by the first author, who holds PhD in psychology as well as a licenced psychotherapist. This manual verification ensured conceptual validity and a balanced class distribution. Subsequently, the dataset was randomly divided into two subsets of 100 items each: one was used as the RAG knowledge base, and the other was for the model evaluation.

2.2. Experiments

Two models were compared in this study.

2.2.1. Baseline (non-RAG) model

We adopted the GPT-4o-mini model to classify each statement independently. No additional context, memory, or external information was provided to the model. The model received only the target sentence along with a fixed classification instruction and was prompted to output a JSON-formatted prediction. The prompt was as follows:

```
You are a psychology content classifier.
[USER STATEMENT]
{user_text}
Classify and return STRICT JSON with the following keys:
- "diagnosis": 1 or 0 (1 means "delusional", 0 means "non-delusional")
- "severity": null or 1/2/3 (1=mild, 2=moderate, 3=severe); only set if diagnosis is "
    delusional"
- "confidence": integer 0-100 indicating classifier confidence (not clinical certainty)
```

```
- "reasoning": one brief, neutral sentence explaining the decision (no PHI, no clinical
    claims)
Return ONLY the JSON.
```

This condition served as a reference point to evaluate how well the LLM could detect delusional content without supplementary information.

2.2.2. RAG model

The second model used the same GPT-4o-mini classifier but incorporated retrieval-augmented generation (RAG) techniques. For each test statement, the system retrieved the three most semantically similar examples from the 100-item retrieval corpus using sentence embeddings and cosine similarity. Sentence embeddings were generated with the text-embedding-3-large model [11], producing fixed-length vector representations of each statement. All embeddings were stored and indexed, and for each test sample we used scikit-learn's `NearestNeighbors` module with cosine distance to identify the top-3 nearest neighbours in the embedding space [12]. This k-NN retrieval step formed the retrieval component of the RAG pipeline. These retrieved examples were concatenated to the prompt as contextual information. The full prompt was:

```
You are a psychology content classifier.
[USER STATEMENT]
{user_text}
[RETRIEVED EXAMPLES] (similar statements from the corpus; may be imperfect)
{ctx_str}
Classify and return STRICT JSON with the following keys:
- "diagnosis": "1" (means "delusional") or "0" (means "non-delusional")
- "severity": null or 1/2/3 (1=mild, 2=moderate, 3=severe); only set if diagnosis is "
    delusional"
- "confidence": integer 0-100 indicating classifier confidence (not clinical certainty)
- "reasoning": one brief, neutral sentence explaining the decision (no PHI, no clinical
    claims)
Return ONLY the JSON.
```

2.3. Evaluation metrics

Because the dataset was balanced and the task involved a binary classification problem (delusional vs. non-delusional), we focused on accuracy, precision, recall, and confusion-matrix analysis as the primary evaluation metrics. Specificity and F1 score were not computed, as the aim of this study was not to provide an exhaustive performance profile, but rather to compare the relative improvement of the RAG model over the non-RAG baseline. These metrics were therefore sufficient to demonstrate the comparative effect of retrieval augmentation.

In addition to classification accuracy, we examined the severity ratings generated by each model. Because severity was coded on a three-level ordinal scale (1 = mild, 2 = moderate, 3 = severe), descriptive statistics (mean, standard deviation, and distribution frequencies) were computed to characterize how each model assigned severity to delusional statements. To assess whether the RAG and non-RAG models differed systematically in the severity they predicted, paired-sample comparisons were conducted using Jamovi, which is a graphical statistical software built on top of the R programming language [13]. For ordinal severity ratings, we ran both nonparametric Wilcoxon signed-rank tests and, for completeness, mean-difference *t*-tests.

2.4. Implementation

All model processing was implemented in Python 3.11 using pandas and scikit-learn libraries. Embeddings were cached to avoid repeated API calls, and each prompt was

Table 1
Confusion matrix for non-RAG model (Accuracy = 0.81).

		Prediction	
		Delusional	Non-delusional
Gold label	Delusional	37	19
	Non-delusional	0	44

Table 2
Confusion matrix for RAG model (Accuracy = 0.94).

		Prediction	
		Delusional	Non-delusional
Gold label	Delusional	50	6
	Non-delusional	0	44

processed with a 50-ms delay to comply with rate-limit requirements. Code reproducibility was ensured through a GitHub repository (https://github.com/YiOliviaHuang/Detection-of-Induced-Psychosis-Like-Thinking-in-User-GenAI-Interaction). Statistical analyses, including descriptive and inferential comparisons of predicted severity against the gold-standard labels, were conducted in Jamovi.

3. Results

3.1. Non-RAG model accuracy

The baseline GPT model achieved an accuracy of 0.81, with a precision of 1.00 and a recall of 0.66. The confusion matrix was shown in Table 1. The model produced no false positives but failed to identify 34% (19 of 56) delusional statements.

3.2. RAG-enhanced model accuracy

In the RAG model, the accuracy was 0.94, with a precision of 1.00 and a recall of 0.89. The confusion matrix (Table 2) indicated that retrieval improved sensitivity while maintaining perfect specificity.

3.3. Comparison of severity: baseline model vs. RAG-enhanced model

Because all non-delusional statements were classified correctly by both models, we examined severity ratings only among statements that were correctly identified as delusional. Both a nonparametric Wilcoxon signed-rank test and a paired *t*-test showed that, compared with the gold-standard labels, the baseline model (Wilcoxon $W = 219$, rank biserial correlation = 0.583, $p = .006$; $t = 3.04$, Cohen's $d = 0.499$, $p = .004$) and the RAG model (Wilcoxon $W = 421$, rank biserial correlation = 0.698, $p < .001$; $t = 4.43$, Cohen's $d = 0.626$, $p < .001$) significantly underestimated severity (see Table 3). Moreover, according to both effect size indicators, the rank biserial correlation and Cohen's d, the baseline model was closer to the gold-standard severity ratings than the RAG model, indicating that RAG further increased the degree of underestimation among correctly diagnosed delusional cases.

4. Discussion

This study examined whether large language models can identify psychosis-like or delusional statements generated during human–AI interactions, and whether retrieval-augmented generation (RAG) enhances

Table 3
Paired samples test.

Measure 1	Measure 2	Test	Statistic	df	p	Effect size	Estimate
Gold label	Non-RAG	Student's t	3.04	36	0.004	Cohen's d	0.499
Gold label	Non-RAG	Wilcoxon W	219^a	–	0.006	Rank biserial correlation	0.583
Gold label	RAG	Student's t	4.43	49	$< .001$	Cohen's d	0.626
Gold label	RAG	Wilcoxon W	421^b	–	$< .001$	Rank biserial correlation	0.698

Note. H_a: $\mu_{\text{Measure 1}-\text{Measure 2}} \neq 0$.
[a] 14 pair(s) of values were tied. [b] 19 pair(s) of values were tied.

this capability. The findings showed that the RAG-based classifier substantially improved diagnostic performance: accuracy increased from 81% to 94%, and recall rose from 0.66 to 0.89 while maintaining perfect precision and specificity. These results indicate that lightweight retrieval augmentation can meaningfully enhance an LLM's sensitivity to delusional content, supporting the view that contextual examples help the model recognise subtle semantic patterns that resemble psychotic thinking.

From a psychological perspective, the ability of an LLM to detect delusional thought patterns is an important step towards developing mental-state-aware safety mechanisms for GenAI systems. Users experiencing psychosis often produce language marked by aberrant attributions, referentiality, and distorted interpretations of agency–features that are difficult to detect without contextual cues. By supplying semantically similar statements, the RAG model appears better able to identify these linguistic markers, suggesting that retrieval-based prompting may help operationalise insights from computational psychiatry within applied AI systems.

However, the severity analysis revealed a different pattern. Both the baseline and RAG models significantly underestimated severity compared with the gold-standard labels. Moreover, the RAG model showed even greater underestimation than the baseline model. This suggests that while retrieval augmentation improves binary diagnostic classification, it does not assist–and may even interfere with–the assignment of severity levels. Severity grading likely requires explicit supervision or clinically grounded reasoning that cannot be inferred simply from semantic similarity. These findings highlight an important divergence between detecting delusional content and evaluating its severity, underscoring the need for more specialised approaches to model-based risk assessment.

Overall, the present study contributes early empirical evidence that RAG can enhance psychosis-like language detection in GenAI systems, while also revealing important limitations in severity estimation that future research should address.

5. Limitations and Future Work

Several limitations should be acknowledged. First, the dataset consisted of simulated statements rather than naturally occurring user–AI dialogues. Although the statements were reviewed and validated by a licensed psychotherapist, real-world data may exhibit greater linguistic variability and complexity. Future studies should incorporate ethically obtained conversational data from actual users, including those experiencing early psychosis or clinical symptoms, to assess ecological validity.

Second, the models were evaluated using prompt-based classification rather than fine-tuned or clinically trained models. While this approach reflects current practice in lightweight safety systems, more advanced techniques–such as supervised training, reinforcement learning from clinical feedback, or hybrid retrieval-and-classification architectures–may yield substantially improved severity estimation.

Third, the severity scale used in this study was coarse (1–3) and relied on single-sentence judgements. Psychosis severity is typically assessed across multiple dimensions (e.g., conviction, preoccupation, distress), which a single utterance may not fully reflect. Expanding severity to multidimensional or continuous measures could improve sensitivity and interpretability.

Finally, this study evaluated only one retrieval method (k-NN with cosine similarity) and one language

model (GPT-4o-mini). Future research should examine alternative embedding models, dense retrieval methods, and larger LLMs to understand the generalizability of the observed effects.

6. Conclusion

This study provides initial evidence that retrieval-augmented prompting can improve the ability of large language models to detect psychosis-like thinking in simulated user–AI interactions. The RAG model demonstrates higher accuracy and substantially better sensitivity than the baseline non-RAG model, suggesting that contextual examples help LLMs recognise delusional content more effectively. However, both models systematically underestimated severity, and the RAG model performed worse than the baseline in this respect, indicating that severity estimation requires more specialised modelling approaches.

Taken together, these findings highlight both the promise and the limitations of using LLMs as early-stage detectors of cognitive risk during human–AI interaction. While retrieval augmentation offers a practical route towards improving safety-sensitive classification, more work is needed to support reliable assessment of symptom severity and to integrate such models into real-world GenAI systems in a clinically responsible manner.

Author Contribution

Yi Huang, Ph.D. in psychology, a licensed psychotherapist, designed the study, prepared and validated the dataset, developed the analysis scripts, and drafted the manuscript.

Data and Code Availability

The code and data were published publicly through https://github.com/YiOliviaHuang/Detection-of-Induced-Psychosis-Like-Thinking-in-User-GenAI-Interaction.

Acknowledgments

No acknowledgments were provided in the submitted manuscript.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT-4o-mini to simulate the initial 200-statement dataset as described in Section 2.1. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the publication's content. Please update this declaration if additional Generative AI tools were used for writing, editing, coding, or figure generation.

References

1. Preda A. Special Report: AI-Induced Psychosis: A New Frontier in Mental Health. American Psychiatric Publishing, Inc. Preprint posted online 2025.
2. Morrin H, Nicholls L, Levin M, et al. Delusions by design? How everyday AIs might be fuelling psychosis and what can be done about it. Preprint posted online 2025.
3. Hill K, Freedman D. Chatbots can go into a delusional spiral: here's how it happens. *New York Times*. 2025.
4. Fieldhouse R. Can AI chatbots trigger psychosis? What the science says. *African Journal of Ecology*. 2023;61:226–227.

5. Ostergaard SD. Will Generative Artificial Intelligence Chatbots Generate Delusions in Individuals Prone to Psychosis? *Schizophrenia Bulletin*. 2023;49(6):1418–1419. doi:10.1093/schbul/sbad128.
6. Hill K. They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling. *The New York Times*. 2025. Available from: https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html.
7. Klee M. People Are Losing Loved Ones to AI-Fueled Spiritual Fantasies. *Rolling Stone*. 2025. Available from: https://www.rollingstone.com/culture/culture-features/ai-spiritual-delusions-destroying-human-relationships-1235330175/.
8. Corcoran CM, Mittal VA, Bearden CE, et al. Language as a biomarker for psychosis: A natural language processing approach. *Schizophrenia Research*. 2020;226:158–166. doi:10.1016/j.schres.2020.04.032.
9. Morgan SE, Diederen K, Vertes PE, et al. Natural Language Processing markers in first episode psychosis and people at clinical high-risk. *Translational Psychiatry*. 2021;11(1). doi:10.1038/s41398-021-01722-y.
10. American Psychiatric Association. Schizophrenia spectrum and other psychotic disorders. In: *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*. 2013:87–122.
11. OpenAI. OpenAI API documentation. 2025. Available from: https://platform.openai.com/docs.
12. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 2011;12:2825–2830.
13. Sahin M, Aybek E. Jamovi: an easy to use statistical software for the social scientists. *International Journal of Assessment Tools in Education*. 2019;6(4):670–692.