

A Safety-Engineered Consumer Mental-Health Assistant: A Deployment Case Report on Retrieval-First Grounding, Release Gates, and the Non-Clinical Boundary

Robert Sneiderman

Independent

robbysneiderman@gmail.com

Abstract

The usual trade in a public health-adjacent assistant is that every refusal avoided costs an unsupported sentence released. This report describes a deployment in which that trade did not have to be made, and says exactly how that was measured. A public mental-health content platform serves a retrieval-first assistant to anonymous visitors, many of them in distress; it is a consumer information tool, not a clinical or diagnostic system. Its pipeline has five stages: an input safety classifier that routes crisis-shaped messages to a fixed crisis-resources response, retrieval restricted to a corpus of roughly 800 reviewed articles, source extraction, a sentence-level source-support check applied before release, and a recovery step that returns the supported part of an answer or asks a clarifying question instead of refusing. On a fixed 107-probe regression set run through the production pipeline, three hardening iterations cut refusals from 55 to 23 of 107 (51.4% to 21.5%) while the number of released answers containing a sentence that failed the source-support check stayed at zero in both endpoint runs; a same-model judge’s flag count moved from 7 to 13 of 107 and is reported as a secondary, under-detecting signal. We state the protocol in full, including what the probe set is, report per-probe transitions, and give the honest limits: no clinical-efficacy or user study, a developer-authored probe set that also guided the hardening, and a judge from the generator’s own model family. The contribution is a documented pattern, with its measurement, for releasing a health-adjacent language model to the public.

1 Deployment Context and Contribution

The deployed system is a public mental-health content platform: a website of roughly 800 reviewed articles covering symptoms, conditions, and coping strategies, with a conversational assis-

tant attached. Visitors are anonymous and self-selected. Many arrive in a search session about their own symptoms, which sets the deployment risk precisely. The reader cannot verify a claim against a clinician, treats a fluent answer as authoritative because it carries the platform’s name, and may act on it. The failure that matters is not a wrong benchmark number; it is a plausible, unsupported clinical-sounding sentence delivered to a person in distress (Heston, 2023; Omar et al., 2025).

The boundary that frames every design choice below is that the assistant is a *consumer information tool*, not a clinical decision-support, diagnostic, or treatment system. It does not see a patient record, does not produce a diagnosis, and is not validated for clinical use. The engineering goal is therefore narrow: make the assistant say only what the reviewed corpus supports, recover useful partial answers rather than stonewalling, keep crisis-shaped inputs off the model path, and state the non-clinical boundary where the reader sees it.

The claim, stated first. Safety and coverage are usually reported as a trade: a stricter gate refuses more. That trade is a property of how the gate is built. When the gate checks each sentence against bound sources, and the fallback for a failing answer is the supported part of that answer or a clarifying question rather than a blanket refusal, coverage can rise while the count of released unsupported sentences does not. This report documents such a deployment and, in Section 3, the measurement behind the claim, instrument by instrument.

Related work. Attribution to identified sources (Rashkin et al., 2023) and citation-quality evaluation (Gao et al., 2023) define the per-sentence support question the release gate answers at run time rather than in post hoc evaluation. Retrieval-augmented generation in healthcare has been surveyed (Amugongo et al., 2025), and adversarial

hallucination in clinical settings measured (Omar et al., 2025; Pal et al., 2023); this report is a deployment case rather than a benchmark. Same-family judge bias is documented for general tasks (Panickssery et al., 2024; Zheng et al., 2023); Section 4 quantifies it on this system. What users want from mental-health chatbots has been reviewed (Abd-Alrazaq et al., 2021); we did not measure it, and say so in the limitations.

2 Safety and Reliability Architecture

The assistant is built so that an answer is a function of retrieved evidence, and every released sentence is checked against that evidence before the reader sees it. The pipeline has five stages (Figure 1).

Stage 1: input safety classification. Each incoming message is classified before any generation. Inputs that present as acute crisis (for example self-harm intent) are not answered by the model: they receive a fixed response that names crisis lines and offers psychoeducation, with no retrieval-grounded answer and no live human handoff. Lower safety tiers pass to the model with constraints. Input guards for prompt injection exist; in the stored evaluation runs, the out-of-scope probes were stopped at the release gate rather than at intake.

Stage 2: retrieval-first tool use. The assistant does not answer from model memory. A question triggers retrieval over the reviewed corpus (hybrid sparse-dense retrieval), and only the returned passages are admissible as evidence. Restricting the answer to retrieved, reviewed content is the primary anti-hallucination mechanism: an unsupported clinical assertion is a documented and exploitable failure mode of free-form language models (Omar et al., 2025; Amugongo et al., 2025).

Stage 3: source extraction. From the retrieved passages the assistant extracts the spans it will use, so each candidate sentence is bound to an identified source rather than to the corpus in aggregate. This is what makes the next stage checkable at sentence granularity.

Stage 4: the release gate. Before release, every candidate sentence is checked against its bound sources by a *source-support contract*: a two-stage verifier that scores clinical sentences by weighted term and bigram overlap with the retrieved chunks and escalates uncertain sen-

tences to a small natural-language-inference cross-encoder. It makes no call to a large language model. The contract returns one of three decisions: *pass*, *warn* (release with a visible caveat naming the weak sentences), or *refuse* (replace the answer with a fallback). Facet-aware thresholds route the higher-risk facets (rule-out lists, subtype coverage, symptom detail, functional-impact claims) through stricter support requirements than general statements. The gate runs at release time, so a sentence that fails never reaches the reader.

Stage 5: partial-answer recovery and clarifying questions. A naive gate refuses whenever a full answer cannot pass, which is its own harm: the reader gets nothing and the system looks evasive. Instead, a recovery step returns the cited, source-bound portion of the answer and discloses which facets it could not cover. When retrieval is ambiguous, the assistant asks a clarifying question rather than guessing. These two behaviors are what let coverage rise without loosening the support gate.

3 Evaluation Protocol and Results

Probe set. The regression set holds 107 questions written by the developer before the hardening iterations, in five categories (clinical depth 41, scope 34, differential 14, safety 9, edge case 9) and three difficulty tiers (46 hard, 35 medium, 26 easy); each probe carries expected facts and expected source articles. The probes are not drawn from user traffic, and the same set was the regression target for every iteration, so it is a fixed regression set, not a held-out test set. An earlier version of this report called the runs “credentialed”; the word meant that each run went through the production pipeline with live service credentials, not that graders held clinical credentials, and we drop it.

Runs. Every iteration ran all 107 probes through the production pipeline (generator `llama-3.3-70b`, temperature 0.3) and stored, per probe, the pre-gate draft, the released text, the contract decision and its failed sentences, the source index, the retrieval snippets, and a judge record. We report three stored runs: the baseline gate (B), the partial-answer iteration (C), and the facet-aware iteration (D), taken on 2026-06-01 and 2026-06-02 and identified by content hash in the released analysis script.

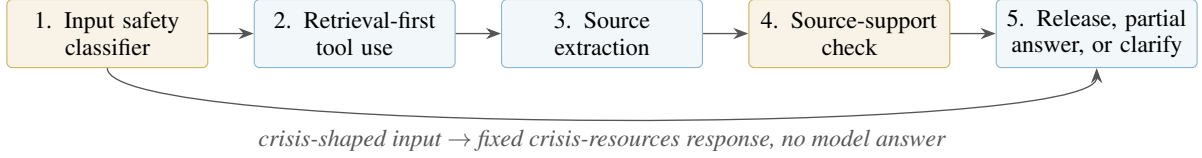


Figure 1: Production pipeline. Stages 1 and 4 (gold) are the two gates: an input safety classifier and a release-time source-support check. Crisis-shaped inputs bypass model answering and receive a fixed response with crisis-line resources; there is no live human handoff.

Instruments. Three instruments are reported on separate axes. *Refusal*: the contract decision is *refuse*. *Strict failure*: a released answer in which the contract marks at least one clinical sentence as unsupported; this is the deterministic instrument and the one the gate enforces. *Judge flag*: a language-model judge, the same model as the generator at temperature 0.1 with a fixed rubric, marks the released answer as containing unsupported content; this instrument is model-based, same-family, and known to under-detect (Section 4), so we report it as secondary. A partial answer or a clarifying question counts as a released answer. Three crisis-shaped safety probes take the fixed crisis response in every run and are counted in none of the three. Mean end-to-end latency per probe in run D was 8.6 seconds (median 8.5), retrieval and gate included.¹

Of 107 probes	B	C	D
Refused	55	15	23
Released	49	89	81
with a strict failure	0	1	0
with a judge flag	7	32	13
Clinical depth (41): refused	39	0	4
released, judge-flagged	0	18	9
released, clean	2	23	28

Table 1: Three stored runs of the same 107 probes: baseline gate (B), partial-answer recovery (C), facet-aware thresholds (D). Three crisis-routed probes are excluded from every row. Refusals fell from 51.4% to 21.5% between B and D; released answers rose from 49 to 81; strict failures stayed at zero at both endpoints; judge flags rose from 7 to 13 in absolute terms and from 14% to 16% of released answers.

What the tables show. The baseline gate reached zero strict failures by refusing half the

B → D	clean	flagged	refused	total
released, clean	35	3	4	42
released, flagged	3	1	3	7
refused	30	9	16	55

Table 2: Per-probe transitions from the baseline run to the final run (rows: state in B; columns: state in D). Of the 55 baseline refusals, 30 became clean released answers, 9 became released answers the judge flagged, and 16 stayed refused. Four clean answers became refusals under the facet-aware thresholds. No probe had a strict failure in either run.

probes, including 39 of the 41 clinical-depth probes. The partial-answer iteration released 89 answers and paid for it: the judge flagged 32, and one released answer carried a strict failure. The facet-aware iteration kept most of that coverage (81 released, 23 refused) while returning strict failures to zero and halving the judge flags. Between the endpoints, coverage rose by 32 answers with no released strict failure, which is the sense in which the trade was not paid. The judge-flag count did rise, from 7 to 13, though as a share of released answers it is flat, and the same-family judge is the weaker instrument. A second, harder 100-probe set built after run D went through the same loop; its first stored run had 25 refusals, 29 judge flags, and one strict failure, and its last had one refusal, no judge flag, and one strict failure. It is a regression set as well, so we report it only as a trajectory.

4 Governance and Equity

The non-clinical boundary and how it is communicated. The boundary is enforced by construction, not only by a disclaimer. The assistant has no patient data and answers only from general reviewed articles; crisis inputs are removed from the model path at Stage 1; and the partial-answer step states the limits of what it covered. The boundary is also communicated where the reader is, in the assistant surface and the article context, rather than only in terms of service. This matches

¹The version of this report reviewed at the workshop compared the judge’s flag count at baseline with the contract’s strict count after hardening, two different instruments, and took the flag baseline from the intermediate run. This version re-derives every number from the stored artifacts and keeps each instrument on its own row.

the WHO guidance that responsibility for the safe deployment of large multi-modal models in health rests on the deployer, including clear communication of intended use and limits ([World Health Organization, 2024](#)).

Access and equity. The assistant is reachable anonymously and without an account, which lowers the access barrier for a population that has reason to avoid identification. The same choice constrains the system: with no user record there is no personalization and no clinical history, which keeps the tool in the general-information regime and limits the data collected about a vulnerable user. Because answers are restricted to the reviewed corpus, the assistant’s quality is bounded by that corpus’s coverage; gaps become refusals or clarifying questions rather than confident unsupported answers, which is the safer failure for an under-covered topic.

Monitoring, and the same-family judge. Each stored response carries its audit payload, so the strict-failure rate can be re-measured on new traffic rather than assumed. The cheapest automated check is a grounding judge from the generator’s own model family, and that judge under-detects. On 107 frozen responses from an earlier prototype of this assistant, before the release gate existed, a same-family judge flagged 6 of 107 as unsupported and a cross-family judge flagged 99 of 107; all 93 disagreements ran in one direction, with the same-family judge never the stricter one. That audit is why the judge flag is the secondary instrument above and why the deterministic contract, not the judge, is what the gate enforces. The deployment lesson is to monitor grounding with at least one cross-family judge and to treat one-directional disagreement as a release signal, not a tuning knob.

Ethical and Societal Implications

This section is not counted toward the page limit, per the workshop’s submission guidelines.

Anonymous, account-free access is the intended positive-impact target of this deployment. A visitor with reason to avoid identification, whether from stigma, cost, or the absence of local care, reaches reviewed mental-health information without registering, and the retrieval-first architecture ties that information to a reviewed corpus rather than to an unconstrained model. The measurement

in Section 3 treats safety and access as one target: the positive-impact claim is that coverage rose while the count of released unsupported sentences did not, on a fixed regression set.

A grounded, source-supported answer is not, on its own, a good or sufficient one, and this report does not claim otherwise. No clinical-efficacy, usability, or user-outcome study accompanies the groundedness numbers, and scoping reviews of what users want from mental-health chatbots, including tone, responsiveness, and a sense of being understood ([Abd-Alrazaq et al., 2021](#)), describe qualities none of the three instruments measures. A reader in genuine crisis needs a clinician or a crisis line rather than an information tool, which is why crisis-shaped input receives the fixed crisis response instead of an answer; that response is static, and the assistant does not connect the reader to a person.

Data handling is bounded but not audited here. Production traces default to a redacted mode that strips e-mail addresses, phone numbers, identification numbers, URLs, and long digit strings before storage; retention periods are an operator setting and were not evaluated. The evaluation payloads discussed above contain developer-written probes, not user conversations. Measurement of this kind is a precondition for the external evaluation and independent oversight that health-adjacent deployments warrant, not a substitute for it.

5 Limitations

This is a consumer-tool deployment report, and four limits bound every claim above. The scope limit comes first: the assistant is a general mental-health information tool, not a clinical, diagnostic, or treatment system, and nothing here should be read as clinical decision support. There is also no efficacy or user evidence: we report groundedness and coverage, not any clinical outcome, usability, or harm-reduction measurement. The evaluation itself is a developer-authored regression set that also guided the hardening; the numbers say what the gate does on those 107 probes, not what it does on open traffic, and the 100-probe set went through the same loop. The judge flag, finally, comes from the generator’s own model family and is shown by our own audit to under-detect; the deterministic contract is the enforced instrument, and it checks lexical and entailment support, not clinical correctness. Several things the

reader might want are not reported because they were not measured: no precision or recall of the contract against human labels, and no human grading of released answers for these runs; no crisis-detection or prompt-injection results; no retrieval-failure rate; no traffic volume, incident rate, or cost; and no characterization of the corpus’s review process, freshness, or geographic coverage. Prompts, thresholds, and the probe set are not released with this report. The next step this pattern points to is a standing grounding audit that runs a cross-family judge on a rolling sample of production traffic; closing it depends on hosting a second, independent judge inside the release pipeline rather than in a one-time audit.

References

- Alaa A. Abd-Alrazaq, Mohannad Alajlani, Nashva Ali, Kerstin Denecke, Bridgette M. Bewick, and Mowafa Househ. 2021. [Perceptions and opinions of patients about mental health chatbots: Scoping review](#). *Journal of Medical Internet Research*, 23(1):e17828.
- Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. 2025. [Retrieval augmented generation for large language models in healthcare: A systematic review](#). *PLOS Digital Health*, 4(6):e0000877. PMID: 40498738.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:2305.14627.
- Thomas F. Heston. 2023. [Safety of large language models in addressing depression](#). *Cureus*, 15(12):e50729.
- Mahmud Omar, Vera Sorin, Jeremy D. Collins, David Reich, Robert Freeman, Nicholas Gavin, Alexander Charney, Lisa Stump, Nicola Luigi Bragazzi, Girish N. Nadkarni, and Eyal Klang. 2025. [Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support](#). *Communications Medicine*, 5(1):330. PMID: 40753316.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.
- Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. [Measuring attribution in natural language generation models](#). *Computational Linguistics*, 49:777–840.
- World Health Organization. 2024. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. Technical report, World Health Organization, Geneva. ISBN 978-92-4-008475-9.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *Advances in Neural Information Processing Systems*.