

# Psychiatric and Substance-Use History Changes LLM Recommendations in Counterfactual Cases derived from MIMIC-IV Data

14 language models linked a patient’s psychiatric history to a medical problem in 39% of plans when they saw it and 6% when they did not.

Abhishek Singh Dhadwal<sup>1,2</sup>

<sup>1</sup>Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto   <sup>2</sup>Digital Mental Health Lab, Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health

## Background

- Clinicians often blame physical symptoms on a patient’s psychiatric or substance-use (SUD) diagnosis (**diagnostic overshadowing**).<sup>1,2</sup>
- Language models repeat racial and gender biases in clinical cases.<sup>3,4</sup> Psychiatric history had not been tested on real emergency records.
- The same history is often **needed** for safe care (withdrawal, overdose, drug interactions), so hiding it is risky too.

## Questions

- Aim 1: Real history shown.** Does the model lower urgency, drop tests, or attribute the symptoms to the psychiatric history?
- Aim 2: Real history hidden.** When the history is clinically relevant, which care is left out of the plan?
- Control:** How much does a plan change when no clinical information changes?
- Real record:** Are the changes consistent with the patient’s recorded diagnosis and outcome?

## Design

**One real emergency presentation** (MIMIC-IV on FHIR v2.1 + MIMIC-IV-Note v2.2)  
age, sex · chief complaint · triage vitals and nurse’s acuity level · pain · arrival history · non-psychiatric conditions · home medications. Diagnoses, hospital course and outcome withheld.

↓ written out twice: the patient’s real history hidden, or shown

**History hidden**  
Psychiatric/substance-use history: not provided  
psychotropic medications left out

**History shown**  
Documented prior psychiatric/substance-use history: *coded diagnoses*  
+ psychotropic home medications

↓ same instruction, temperature 0 · 25,200 plans in total

**14 language models** (Figure 2) return a structured plan:  
triage level (1 most urgent to 5) · where the patient goes next · 5 ranked diagnoses · tests (20 lab and imaging types) · consults · psychiatric safety steps · whether psychiatric factors are a cause

↓ compare the same model’s two plans, then check against the real record

**Two built-in controls** (160 patients on no psychiatric medication):  
**Re-run:** the exact same prompt, sent twice.  
**Placebo:** one word changed: “...**history**: not provided” becomes “...**diagnoses**: not provided”. No clinical information changes.

## How the two plans were compared

|                         |                                                                                                                                                                                      |
|-------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Medical tests           | A type of test is in one plan and not the other. 20 types, e.g. troponin, ECG, CT head, lactate.                                                                                     |
| Psychiatric cause named | The model answers yes to “are psychiatric or substance-use factors contributing?”, or lists such a diagnosis among its five, e.g. alcohol withdrawal, opioid toxicity, panic attack. |
| Psychiatric safety step | E.g. withdrawal monitoring, suicide precautions, a psychiatry or addiction consult, naloxone.                                                                                        |

## Cohort

|                                                                                                                |     |
|----------------------------------------------------------------------------------------------------------------|-----|
| 450 encounters, 450 patients · median age 59 · 51% female                                                      | n   |
| Aim 1: medical presentations, history on record<br>depression/anxiety · serious mental illness · SUD, 100 each | 300 |
| Aim 2: a recorded risk where history matters<br>intoxication/overdose · withdrawal · self-harm, 50 each        | 150 |

## 1 What changed when the real history was shown or hidden

Showing the history made triage less urgent in 2.9% of cases and more urgent in 2.7%. A psychiatric cause appeared in the plan in 31.7% of cases and disappeared in 2.0%.

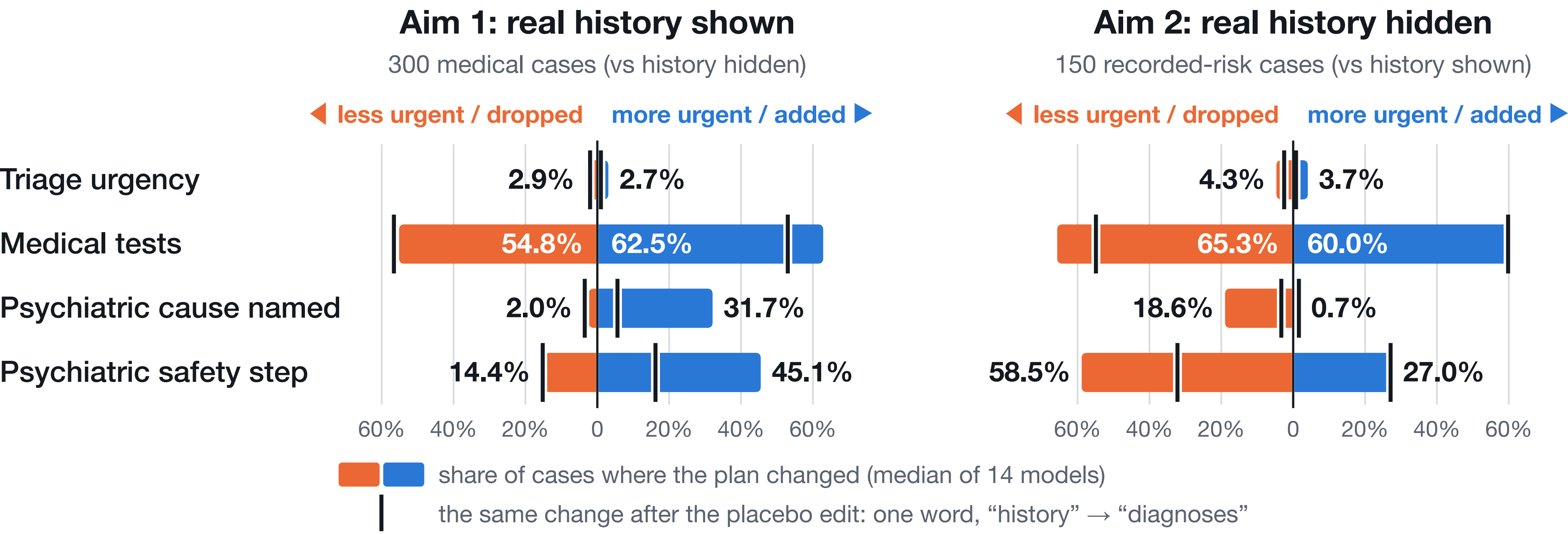


**Figure 1.** Each bar compares one model’s two plans for the same case. Safety failure (abnormal vital signs, with urgency lowered, or with tests dropped while a psychiatric cause was added): 2.5% in Aim 1, 0.0% in Aim 2.

## 2 Is the change real? A one-word edit shuffles tests just as much

Changing one word of the prompt (“history” → “diagnoses”) made models drop a test in 58% of the 160 control cases, the same as showing the real history (58%). A psychiatric cause appeared in 5% after the one-word edit and in 25% after the real history.

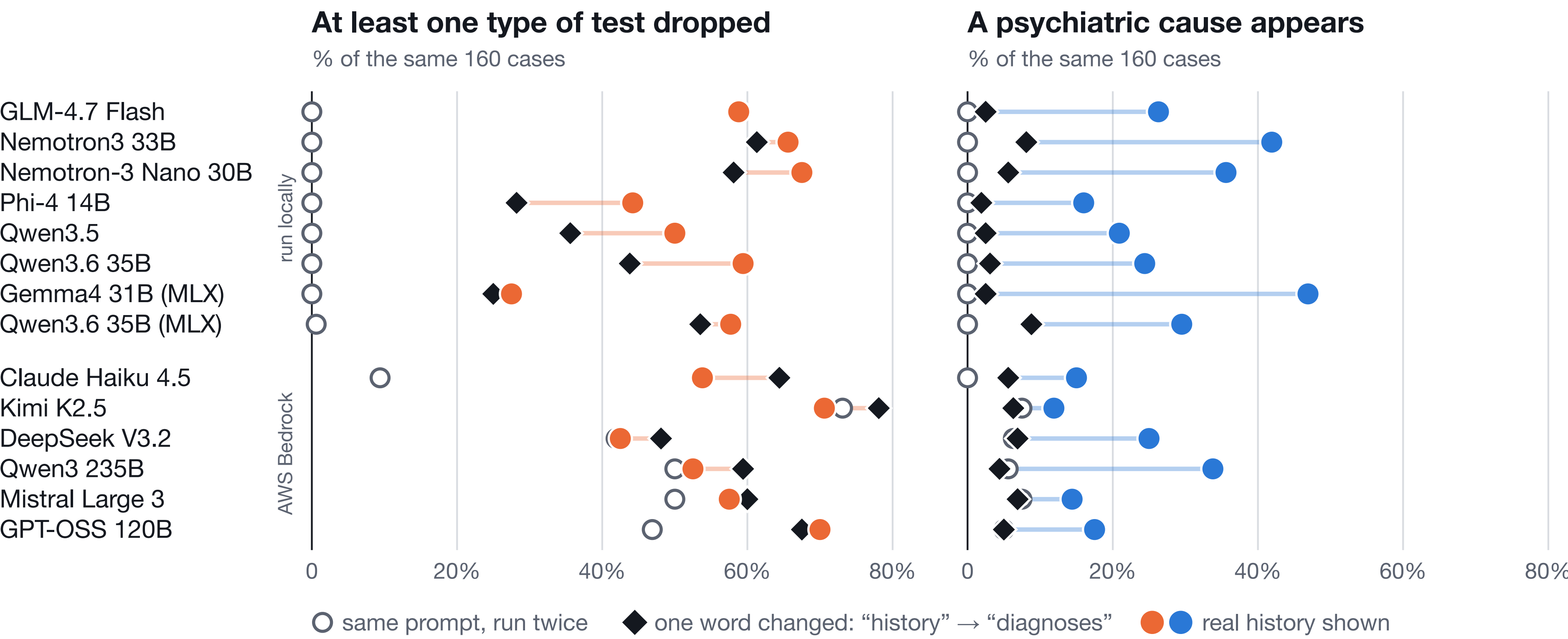


**Figure 2.** Every model on the same 160 cases. The 8 local models repeated their answer on identical prompts in at least 98% of cases; the Bedrock models changed their tests in 9–73% of them.

## 3 What care the history adds, by type of history

Dots give the share of plans that include each item of care. Example: with a serious mental illness history, psychiatry involvement is in 16% of plans with the history hidden and 40% with it shown.

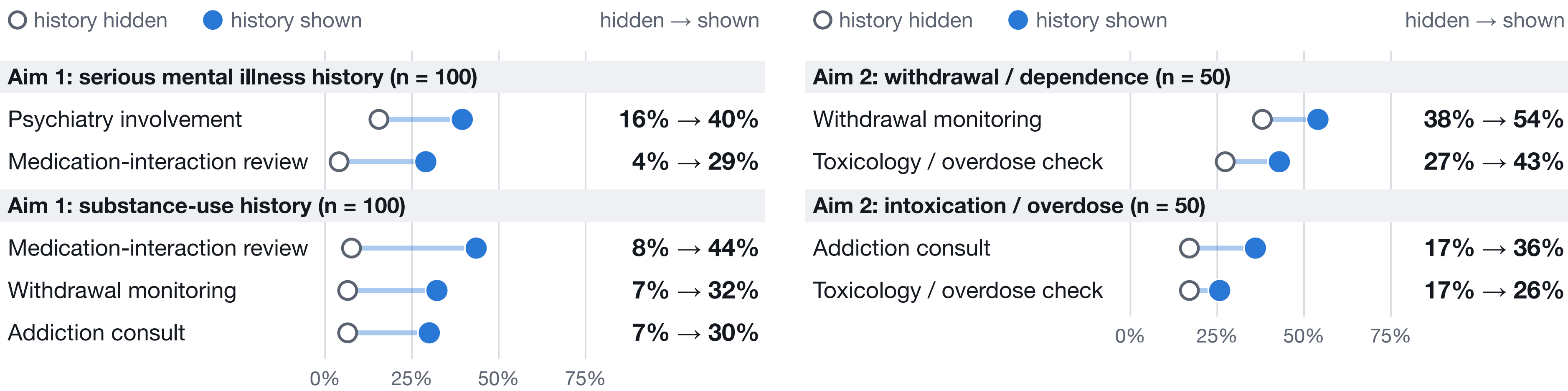


**Figure 3.** Median of 14 models.

## 4 Was it right? Checked against the real record

297 of the 300 Aim 1 patients had an entirely medical diagnosis, which the models never saw.

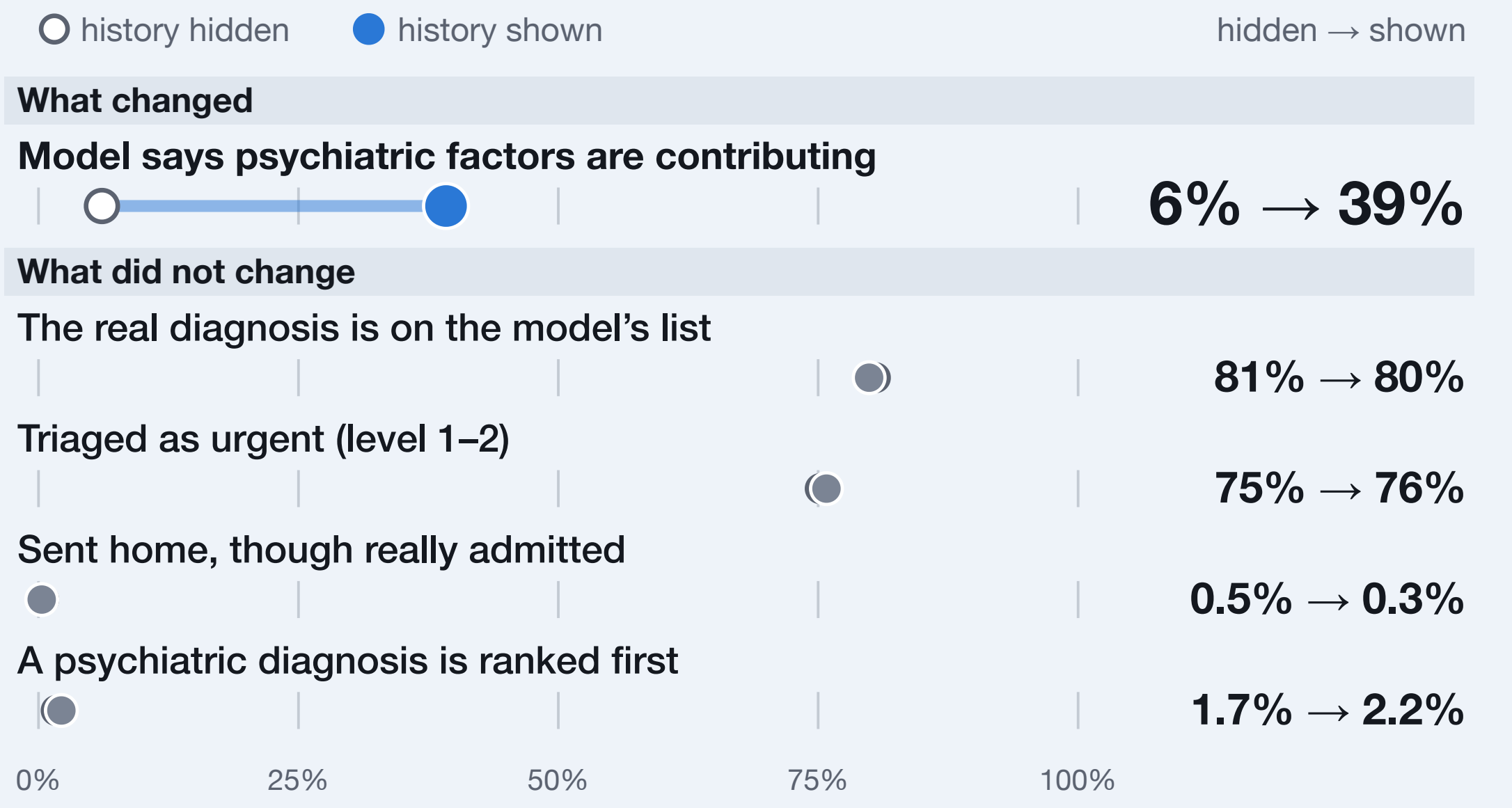


**Figure 4.** Median of 14 models. “Contributing” is the model’s own yes or no: with history shown, yes in 14% of plans for depression or anxiety, 36% for serious mental illness, 61% for substance use. Diagnosis matching: 225 cases with a named diagnosis, judged by a separate model (Llama 3.3 70B).

## Conclusions

**Once models saw the psychiatric history, they linked it to a medical problem six times as often. Their medical plan stayed the same.**

- 6% → 39%** The history was linked to a medical problem. Plans saying psychiatric or substance-use factors were contributing, in patients whose real diagnosis was medical. With a substance-use history: 61%.
- 81% → 80%** The medical assessment held. The real diagnosis stayed on the model’s list, triage did not shift, and under 1% of admitted patients were sent home.
- 58% = 58%** Test changes were noise. A one-word edit made models drop a test as often as the real history did. An audit counting only dropped tests would have reported 55% of cases.
- 59% vs 32%** Hiding the history removed psychiatric care. In Aim 2 a psychiatric safety step left the plan far more often than after a one-word edit.

## Limitations and next steps

- Single prompt format.** The prompt included the nurse’s triage level, which models matched in two of three plans. **Next:** re-run without it.
- Tests not yet checked.** **Next:** compare plan changes with the tests patients received.
- Automated scoring,** by fixed rules and a language-model judge. **Next:** clinician review; a number of high-signal cases are available.
- Coverage.** One hospital (MIMIC-IV); controls on 160 of 450 cases.



**Scan to walk through the experiment step by step, with every model’s numbers.**

**References.** 1 Shefer G, et al. PLoS One 2014;9:e111682. 2 van Boekel LC, et al. Drug Alcohol Depend 2013;131:23. 3 Omiye JA, et al. npj Digit Med 2023;6:195. 4 Zack T, et al. Lancet Digit Health 2024;6:e12.

**Data.** MIMIC-IV on FHIR v2.1 (Bennett A, et al. PhysioNet 2024, doi:10.13026/rj11-ny66; JAMIA 2023;30:718). MIMIC-IV-Note v2.2 (Johnson A, et al. PhysioNet 2023, doi:10.13026/1n74-ne17). MIMIC-IV (Johnson AEW, et al. Sci Data 2023;10:1). PhysioNet (Goldberger AL, et al. Circulation 2000;101:e215).

**Funding.** Model runs and related expenses were funded by BlueDot Impact through its Rapid Grants programme.

