holmusk

# Medication Management LLM

From Free Text to Structured Medication Profiles

Extracting Medication Information from Millions of Neuropsychiatric Clinical Notes

September 6, 2026

## Executive summary

In neuropsychiatry, the fullest picture of what a patient is actually taking rarely lives in a structured database field. It lives in the free-text notes that psychiatrists, nurses, and pharmacists write every day, details on drug, dose, route, timing, and why a treatment changed, described in plain language in clinical notes rather than coded. Even where a structured medication field does exist, both the published literature and Holmusk's own data show it is an incomplete stand-in for what actually happened clinically.

MedLLM (Medication Management LLM) is a purpose built AI model that reads those notes and turns them into structured medication records that can be counted, grouped, and analysed. It has been deployed on a few million de-identified psychiatric notes, a subset of the neuropsychiatric clinical notes generated across more than 10,800 care sites, covering a patient population of over six million. For each medication a clinician mentions, it captures the drug and its dose, route, and frequency, along with any side effects, whether the treatment changed, and the reason behind that change.

This paper is written for partners and prospective clients who want to understand what that capability makes possible, rather than how the model is built internally. This sets out the clinical problem MedLLM was built to solve, what the extracted data looks like in practice, the quality-control pipeline behind it, the patterns already visible in the data, and the safeguards that keep clinicians in control. In keeping with Holmusk's evidence-first approach, the figures here are shared as proportions rather than raw counts, and the full validation study with precision, recall, and expert-adjudicated accuracy is being prepared for peer-reviewed publication.

**What to take away**

- The most useful medication information in psychiatric care is written as free text, and MedLLM makes it usable at population scale.
- It captures not only which drugs patients are on, but how treatment changes over time and why, which is difficult to get from structured EHRs
- MedLLM operates under a responsible AI governance framework: clinical experts remain in the loop at every stage of the process, and all notes are fully de-identified before the model has access to them.

## The documentation gap

In most fields of medicine, the medication list is a fair summary of what a patient is taking. In neuropsychiatry it is often the least informative part of the record. Neuropsychiatric treatment is unusually fluid: doses are titrated up and down, one drug is swapped for another, medications are combined, and drugs are stopped when side effects outweigh the benefit. Almost all of that activity is described in clinical notes, not recorded or coded in a structured field that frequently.

The information itself sits inside the note as ordinary prose: drug name, dose, route, frequency, duration, side effects, and how the regimen changed over time. Notes are written quickly, under time pressure, full of the abbreviations and shorthand that a colleague reads without effort and a database cannot. Mental health care adds further difficulty: much of it has no lab value or objective marker to anchor it, and standardised rating scales are applied inconsistently. The result is a clinical record that is easy for a person to read and very hard for software to use (Kalf et al., 2022; Chapman et al., 2021).

This is not a new observation. Researchers have been building tools to pull clinical facts out of free text for more than a decade, from early language models to today's large ones. What has changed is that the reading

can now be done reliably enough, and at low enough cost, to run across an entire note stream instead of a hand-picked sample.

## Why structured EHR data isn't enough

If an electronic health record already carries a medication list, it's reasonable to wonder why the notes are worth reading at all. They are worth reading because the structured list often tells only part of the story and studies of neuropsychiatric records confirm this. These fields are often left incomplete, and medication lists sometimes still show drugs a patient has already stopped, appearing active in the record when they are not (Kadra et al., 2015; Owen et al., 2011). The notes are what close these gaps: they carry the reason a drug was discontinued, the side effect that drove a switch, and the dosing instruction a patient was actually given.

For anyone using this data for research or real-world evidence, that difference is the whole point. A clean count of prescriptions tells you what was ordered. The notes tell you what happened.

## What MedLLM built

MedLLM is a task-specific model built on MedGemma (Sellergren et al., 2025), Google's open family of medical AI models, fine-tuned for one job: finding medication information in neuropsychiatric notes and writing it out in a consistent structure. Rather than pulling out isolated drug names, it places each medication into the patient's timeline and records the attributes around it -dose, route, frequency, duration, side effects, and, importantly, whether the treatment changed and why. It looks for specific events such as a drug being stopped, a switch to another drug, or a dose being raised or lowered.

The figure below shows the idea in a single illustrative sentence. A short passage of narrative becomes two clean medication records, each with its own attributes and, for the olanzapine, the side effects and the reason it was stopped.

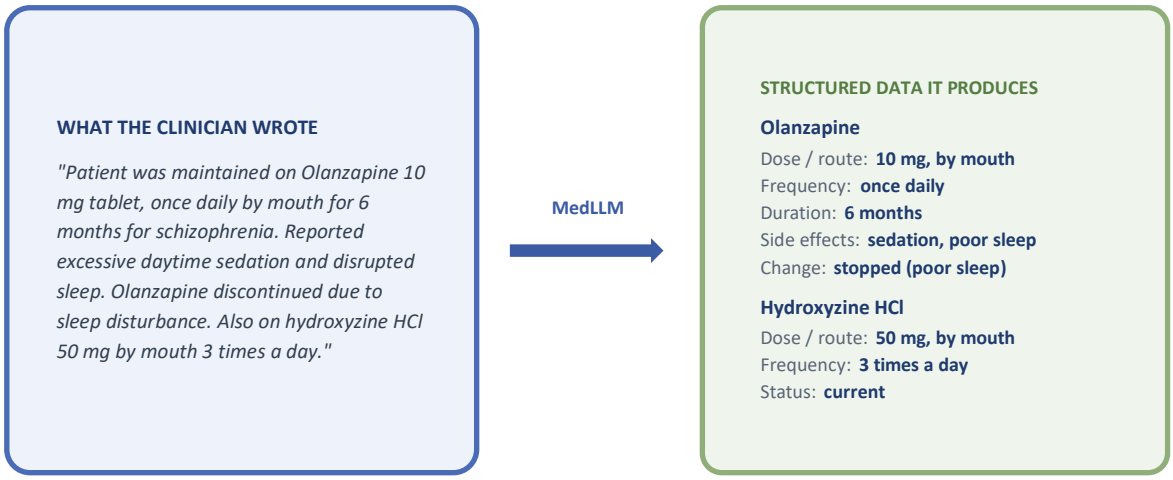


*Figure 1. A worked example on illustrative source text and the resulting structured records.*

The output is structured medication data that researchers and clinicians can query and aggregate. This approach aligns with recent independent work showing that large language models can perform this kind of medication extraction and discontinuation detection at production scale (Shao et al., 2025).

Behind that output is a pipeline built for volume and for scrutiny. Raw notes are de-identified and prepared, the model reads them and proposes structured records, those records pass through layered quality control

that pairs automated checks with clinician review, and only then do they become the structured medication data used downstream.

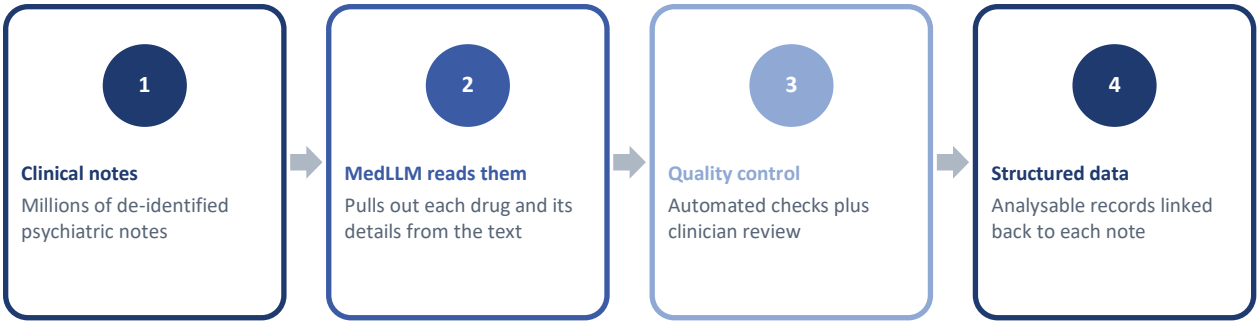


*Figure 2. From raw note to analysable, quality-checked medication data.*

The model was fine-tuned on a large body of neuropsychiatric notes annotated by clinical experts, so it is tuned to how mental-health care is actually documented rather than to generic medical text. And its accuracy is assessed by combining automated evaluation with systematic clinician review rather than relying on either alone, in line with emerging best practice for evaluating clinical language models (Kocaman et al., 2025). This follows the same clinician grounded approach Holmusk has used in its earlier work generating real-world evidence from mental health records (Kulkarni et al., 2024).

## A multi-stage, statistically grounded QC pipeline

Instead of one validation pass, MedLLM's quality-control process runs each output through four checks. Each check is deeper than the last, but covers less ground, starting with an automated check on 100% of the output and ending with a review by clinical experts. The results from each stage feed into a final decision for each entity. This design was updated after the pilot phase to work within real limits: there are only so many expert reviewers, and the system processes millions of notes. It checks all target fields, with extra attention on the ones that are hardest to get right, such as treatment changes and the reasons behind them, since these are the most likely to contain errors.

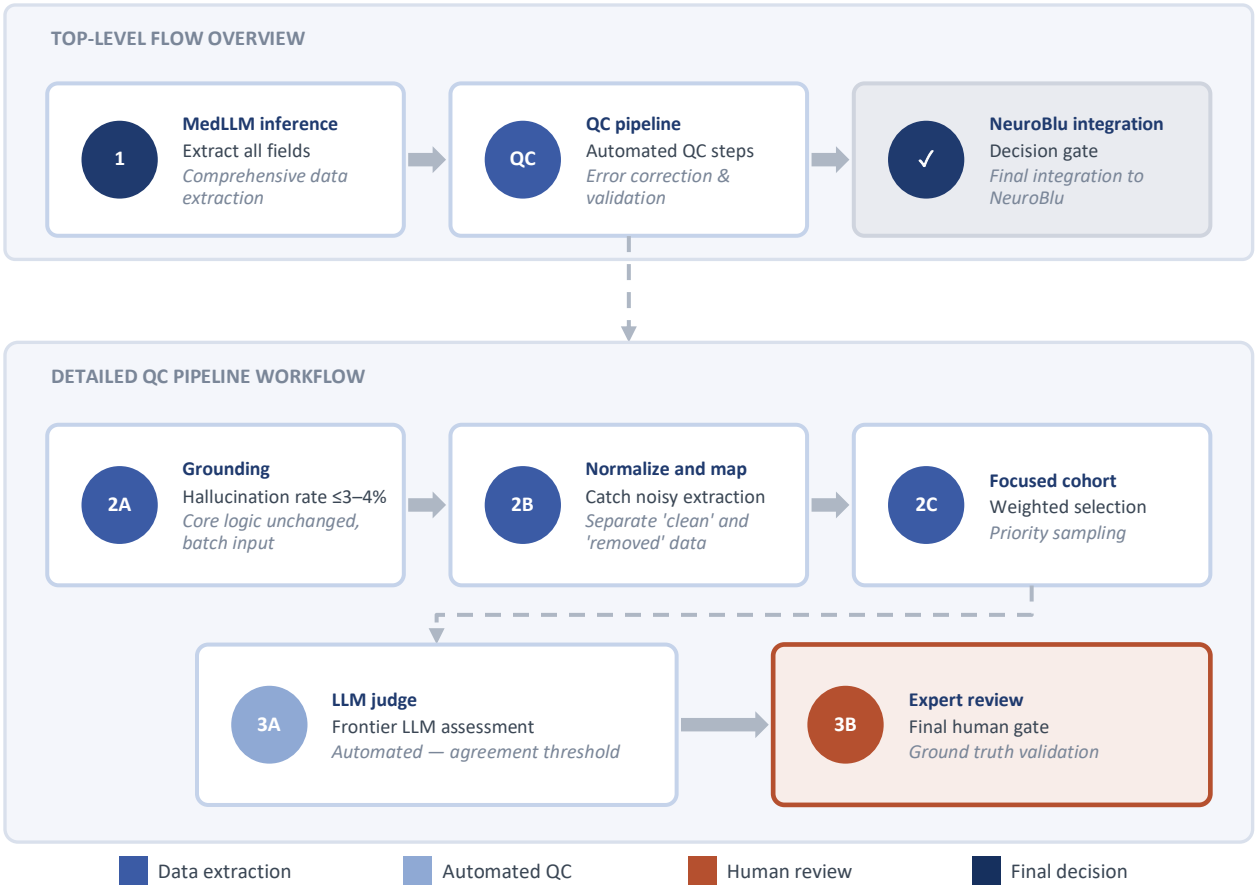


*Figure 3. MedLLM inference QC pipeline, from generation to NeuroBlu integration.*

## Operating at scale, and who the data represents

MedLLM has completed two full production phases, together processing millions of notes. This is an operational milestone that the program has moved from development to sustained deployment across a defined, real-world volume of neuropsychiatric documentation spanning hundreds of thousands of patients and millions of notes. The overall extraction stayed stable across both phases, a reassuring sign that the model behaviour is consistent rather than drifting.

Because the value of the data depends on who it represents, the population is worth looking at directly. MedLLM's output is derived from NeuroBlu, Holmusk's neuropsychiatry real-world data platform. The profiled population, everyone with at least one clinical note, covers over six million patients.

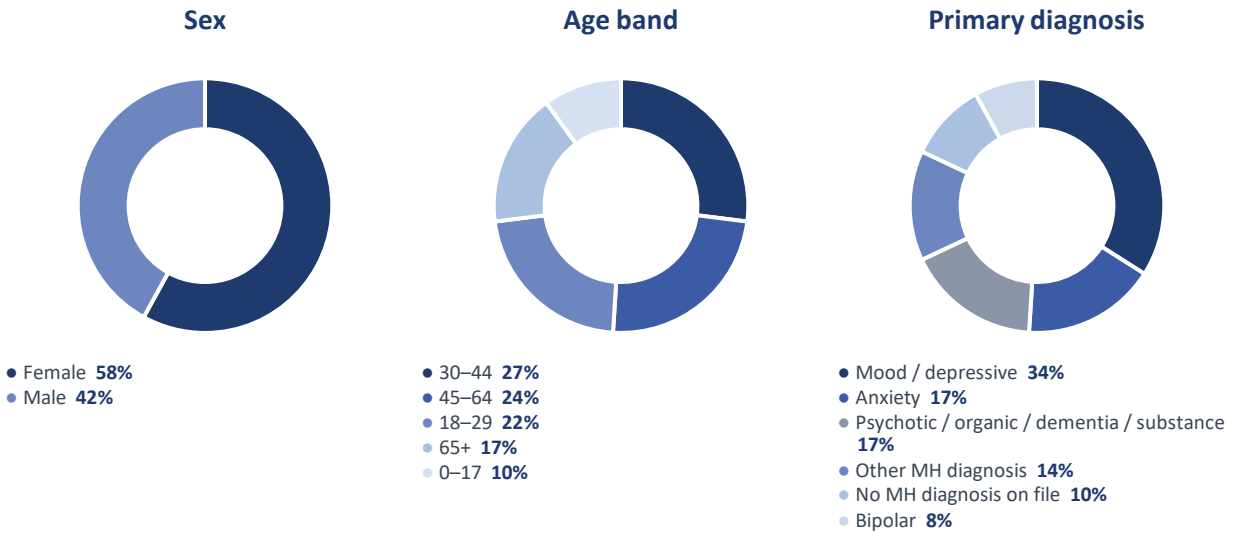


*Figure 4. Profiled population by sex, age band, and primary diagnosis (over six million patients with at least one note).*

The population skews slightly female and sits mostly in working age, and the diagnosis mix is what you would expect for neuropsychiatry care. Mood and depressive disorders are the largest single group, followed by anxiety, with bipolar and the more severe psychotic, organic, and substance-related conditions each making up a meaningful share.

The note stream itself is larger and less even than the patient count suggests. There are roughly 190 million+ notes on file, about 30 per patient on average, across 25 distinct note types.

*(Numbers as of NeuroBlu release 26R3; the count will continue to grow as additional notes are sourced)*

| ~190M | ~30 | 25 |
|---|---|---|
| notes on file across the profiled population | notes per patient, on average | distinct note types in use |

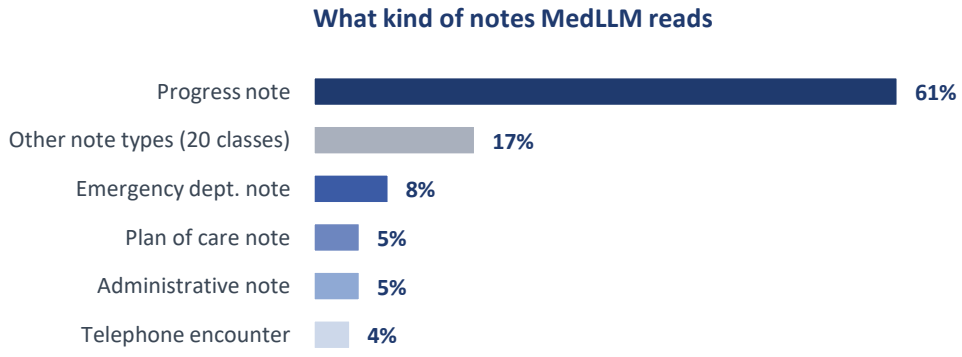


*Figure 5. Progress notes make up roughly three in five notes, with 20 further note types sharing the rest.*

The note mix matters because different note types carry different information. Progress notes are also one of the richest sources of side-effect and treatment change detail, so what MedLLM can extract depends on the kinds of notes it sees, not only how many. Because documentation volume and style vary across more than 10,800 care sites and several EHR platforms, the model is designed to run across the whole stream rather than a curated subset.

## What this unlocks

The value here is not the raw entity counts. It is that a large, stable, structured stream lets researchers and clinicians ask questions that a manual chart review or a plain prescription list cannot answer. Several patterns are already clear, and each points to a concrete use.

### How complete each record is

Across the mentions extracted so far, about three in four carry a dose (74%), and roughly three in five carry a route (59%) or a frequency (58%) , a genuinely complete structured record for most mentions. Duration is the rarest field by far, appearing in only about 2–3% of mentions. The model performs exactly as expected on more contextual fields, which are inherently sparsely documented , changed drug status appears in about 1 in 10 mentions (10%), and side effects and change reasons each appear in roughly 7–8%. These lower numbers reflect real-world clinical documentation patterns, not a limitation of the model's extraction capability. A blank in any of these fields is best read as not documented in this note, not as a confirmed absence.

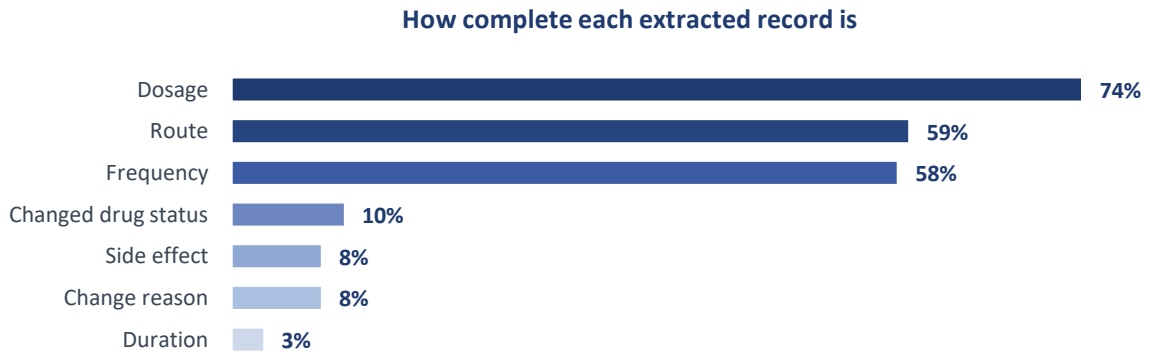


*Figure 6. Completeness of the most common attributes, as a share of extracted drug mentions. These figures reflect the phases completed to date and are likely to shift as the program scales.*

### The pattern of treatment change and the reasons

Because MedLLM captures how treatment changes and not only what was prescribed, it produces something rarely available at this scale: a structured record of drugs being stopped, switched, and adjusted. Among the changes it flags, roughly two-thirds are a treatment being discontinued, about one in seven is a dose increase, and smaller but steady shares are switches between drugs and dose decreases.

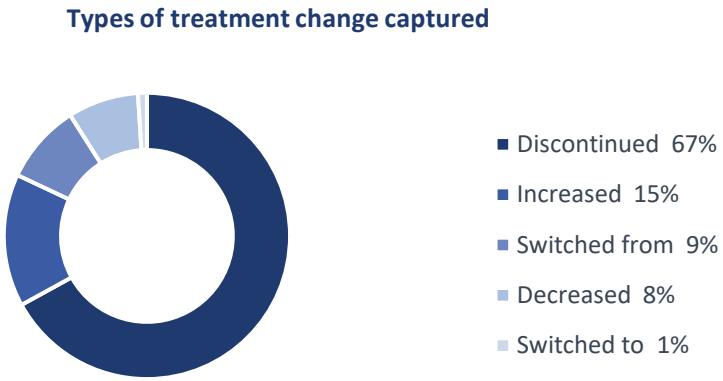


*Figure 7. Treatment-change events, consistent across both scale-up phases.*

A stream of discontinuation and switch events at this volume is exactly what comparative effectiveness and treatment pathway research needs, and it directly fills the medication documentation gap described earlier. At production scale it supports drug switching and discontinuation cohorts, adherence and persistence studies, and research into how medications are combined and augmented: work that would otherwise require reading charts manually.

The change reason text itself tells a consistent story across both scale up phases: lack of treatment response recorded variously as “ineffective,” “not effective,” and “lack of efficacy” is by far the most common reason cited for a medication change, followed by tolerability concerns such as weight gain, generic “side effects,” rash, and nausea.

## A side-effect signal that looks right

MedLLM also surfaces side effects as patients and clinicians describe them. Among the most frequently mentioned, weight gain alone accounts for about one in four, followed by nausea and sedation.

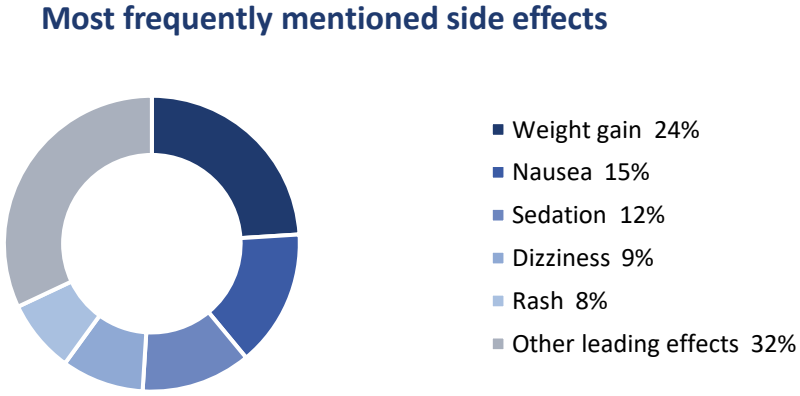


*Figure 8. Most frequently mentioned side effects, consistent across both scale-up phases.*

That ordering is a useful sign that the extraction is capturing something real. Metabolic effects, and weight gain in particular, are among the best-documented burdens of neuropsychiatric medication in the clinical literature, so seeing them rise to the top is what a trustworthy signal should look like (Solmi et al., 2020). A

structured, note derived side-effect stream at this scale could support pharmacovigilance work and metabolic monitoring quality improvement.

### From entities to downstream use cases

Taken together, these patterns and the structural gaps in EHR data described earlier, point toward concrete applications for the structured medication data MedLLM produces:

- **Real-world evidence generation** on psychotropic medication utilization patterns at a population scale not previously accessible from free text alone.
- **Drug-switching and discontinuation cohorts** for comparative effectiveness and treatment pathway research.
- **Pharmacovigilance side-effect surveillance,** prioritizing known high-burden categories such as metabolic and gastrointestinal effects.
- **Quality improvement dashboards** that reconcile a patient's narrative-derived active medication list against the structured EHR record, surfacing discrepancies in active medications documentation, for clinician review.
- **Cohort definition support** for downstream research studies that need medication exposure, switch, or discontinuation status as an inclusion or stratification variable.

## Responsible AI and governance

Because this data comes from sensitive clinical records and is meant to inform research and care, the safeguards around it matter as much as the capability itself. The program is grounded in a responsible-AI framework built around three principles:

- **Rigorous de-identification as a structural prerequisite.** Clinical notes undergo comprehensive de-identification before entering the MedLLM extraction stream. This process involves NER (Named Entity Recognition) to mark identifying fields in the notes and have them be replaced by plausible alternatives. This allows for false negatives to Hide In Plain Sight (Carell D et al., 2013). This process has been audited separately through an Expert Determination process and has been found to meet the standards of de identification set out by HIPAA.
- **Multi-layered quality assurance.** Structured checks and independent review run at every stage of the pipeline, rather than a single check applied at the end.
- **Experts in the loop.** Clinician experts stay in the workflow reviewing and adjudicating output rather than being replaced by it, and remain involved in annotation, review, and governance of how extracted information is generated and used. The intent behind these controls is straightforward: the system is built to support clinical judgement, not to substitute for it. This posture matches emerging responsible AI frameworks written specifically for neuropsychiatric practice, which treat ongoing clinician oversight as a condition for using AI safely in this setting, and it is mindful of a regulatory landscape for AI in mental health that is moving quickly at the state and national level (Mendlovic et al., 2026; Schoene et al., 2026; Governing AI in Mental Health, 2025).

## The evidence roadmap

This white paper describes capability and process, not validation outcomes. MedLLM's full extraction methodology and the QC design detailed above will be published shortly as a peer-reviewed validation study reporting complete performance metrics including cohort statistics, precision and recall, and clinician-

adjudicated concordance for submission to a clinical informatics / NLP journal. This staged publication approach is consistent with the program's commitment to open and reproducible science.

## AI use disclosure

Generative AI tools were used to assist with drafting and graphics for this paper. The core methodology, data, and conclusions were reviewed by human experts before publication.

## References


1. Bui DDA, Zeng-Treitler Q, et al. Classification of medication status change in clinical narratives. PMC. 2011. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041444/

2. Carrell D, et al. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. J Am Med Inform Assoc. 2013 Mar-Apr;20(2) doi: 10.1136/amiajnl-2012-001034 .https://pubmed.ncbi.nlm.nih.gov/22771529/

3. Chapman AB, et al. Extracting information from clinical text: challenges of abbreviation and shorthand. JMIR Medical Informatics. 2021;9(5):e24678. https://medinform.jmir.org/2021/5/e24678

4. Governing AI in mental health: a 50-state legislative review. JMIR Mental Health. 2025. https://mental.jmir.org/2025/1/e80739

5. Kadra G, et al. Extracting antipsychotic polypharmacy data from electronic health records: developing and evaluating a novel process. BMC Psychiatry. 2015;15:166. https://pmc.ncbi.nlm.nih.gov/articles/PMC4511263/

6. Kalf RRJ, et al. Information extraction from free text for aiding transdiagnostic psychiatry: constructing NLP pipelines tailored to clinicians' needs. BMC Psychiatry. 2022;22:762. https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-022-04058-z

7. Kocaman V, Kaya MA, Feier AM, Talby D. Clinical Large Language Model Evaluation by Expert Review (CLEVER): framework development and validation. JMIR AI. 2025. https://ai.jmir.org/2025/1/e72153

8. Kulkarni D, et al. Real-world evidence from mental-health records using clinician-grounded NLP. 2024. https://www.sciencedirect.com/science/article/pii/S2949719123000420

9. Mendlovic S, Frankova I, Vermetten E, et al. Responsible artificial intelligence integration framework for psychiatric guidelines. International Journal of Neuropsychopharmacology. 2026;29(4):pyag010. https://academic.oup.com/ijnp/article/29/4/pyag010/8661708

10. Owen MC, Chang NM, Chong DH, Vawdrey DK. Evaluation of medication list completeness, safety, and annotations. AMIA Annual Symposium Proceedings. 2011. https://pmc.ncbi.nlm.nih.gov/articles/PMC3243276/

11. Schoene AM, Mestre R, Middleton SE, et al. Responsible AI in mental healthcare: policy directions and stakeholder insights. Frontiers in Public Health. 2026;14:1814039. https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2026.1814039/full

12. Sellergren A, et al. MedGemma Technical Report. Google Research & Google DeepMind. arXiv:2507.05201. 2025. https://arxiv.org/abs/2507.05201

13. Shao C, Snyder D, Li C, Gu B, et al. Scalable medication extraction and discontinuation identification from EHRs using large language models. arXiv:2506.11137. 2025. https://arxiv.org/abs/2506.11137

14. Solmi M, et al. Update on weight-gain caused by antipsychotics: a systematic review and meta-analysis. 2020. https://pubmed.ncbi.nlm.nih.gov/31952459/