Catcher Technology Company · Boston Neuromind LLC

Conversational Differential Diagnosis Loop: An Integration Algorithm for Verifiable Multi-Component Psychiatric Triage in Dialogue

Alex[1]

[1] *Boston Neuromind LLC, Boston, MA, USA. Former Visiting Scholar, Harvard Graduate School of Education.*

Correspondence: Boston Neuromind LLC, Canton, MA, USA.

Abstract

Background. Mental health chatbots have moved rapidly from rule-based systems to large language models (LLMs), yet a recent systematic review of 160 studies found that only a small fraction of LLM-based systems have undergone clinical efficacy testing, and most remain in early validation. A central unmet need is a triage agent that can guide a user who does not yet know "where it hurts" toward an appropriate area of clinical concern — without issuing a diagnosis, and through a conversation that feels comfortable rather than like an interrogation. Objective. We specify the Conversational Differential Diagnosis Loop (CDDL), an integration algorithm that combines three established research traditions — adaptive measurement, transdiagnostic dimensional structure, and Bayesian sequential differential reasoning — inside the medium of natural dialogue. Method. CDDL is built around a single verifiable objective: at every turn the agent selects the question that maximizes expected information gain over a live hypothesis board. We decompose this objective into four terms (DiagnosticSeparation, TemporalValue, DimensionalCoverage, ConversationalCost), give the mathematical form of each, and separate an immutable safety "constitution" (red-flag scanning, DSM gate questions, the not-a-substitute boundary) from an evolvable "phenotype" (dialogue style, curated weights). Evolution proceeds by curated version increments under a locked-algorithm principle, not by self-adjustment. Results. We show that the integration produces three properties that no individual component exhibits alone: temporal-aware question selection, confidence-modulated conversation, and information-gain stopping. We provide the per-turn computation, a worked dialogue trace, simulation of entropy reduction against fixed-questionnaire and random-order baselines, and a validation protocol expressed as predicted-versus-observed entropy reduction. Conclusion. CDDL reframes the contribution of a conversational triage agent: the components are established; the integration is the contribution — and, because it embeds curation and clinical judgment inside the loop, the defensible asset.

Keywords: conversational agent; differential diagnosis; psychiatric triage; adaptive testing; transdiagnostic; HiTOP; information gain; measurement-based care; large language models; digital mental health.

Contents

Contents....3
1. Introduction....5
1.1 The problem: a user who does not know where it hurts....5
1.2 Two questions, asked at 2 a.m....5
1.3 The gap, confirmed by review....5
1.4 Contributions of this paper....6
2. Background and Related Work....7
2.1 Adaptive measurement — solid ground for question selection....7
2.2 Transdiagnostic dimensional structure — solid ground for the dimension layer....7
2.3 Bayesian sequential differential reasoning — solid ground for the board and selector....7
2.4 Conversational LLM mental health assessment — the thin ground....8
2.5 Longitudinal and measurement-based care — solid ground for the within-user axis....9
3. The CDDL Architecture....11
3.1 Seven design principles....11
3.2 The four components of the triage engine....11
3.3 The loop — one turn....12
4. The CDDL Function....13
4.1 The heart of the loop: the next-question selection function....13
4.2 Term 1 — DiagnosticSeparation....14
4.2.1 Resolving the likelihood from a descriptive vault....14
4.3 Term 2 — TemporalValue....15
4.3.1 The within-user store and trait / state / course....16
4.4 Term 3 — DimensionalCoverage....17
4.5 Term 4 — ConversationalCost....17
4.5.1 The sensitivity tag and the constitutional exception....18
5. Weights, the Per-Turn Computation, and a Worked Trace....19
5.1 The weights w_1 w_2 w_3 w_4 — handling “it differs by person”....19
5.2 The per-turn computation....19
5.3 A worked dialogue trace....20
6. Constitution, Phenotype, and Curated Evolution....22
6.1 Why self-adjusting learning is excluded....22
6.2 Curated version evolution....22
6.3 The constitution / phenotype separation....22
7. Three Emergent Properties of the Integrated Loop....24
7.1 Temporal-aware question selection....24
7.2 Confidence-modulated conversation....24
7.2.1 Formalizing confidence modulation in the function....24
7.3 Information-gain stopping....25
8. Validation Protocol....27
8.1 Term-level validation....27
8.2 Loop-level validation....27
8.3 What this paper does and does not claim....28
9. Scope, Limitations, and Future Work....30
9.1 What v1 excludes....30
9.2 Limitations....30
9.3 Future work....30
10. Conclusion....31
References....32

1. Introduction

1.1 The problem: a user who does not know where it hurts

A person who is unwell physically can usually name a region of the body. A person who is unwell psychologically often cannot. They arrive with a diffuse report — “I feel off,” “nothing is fun anymore,” “I can’t focus” — that is compatible with many diagnostic possibilities at once. The clinical task that follows is not to confirm a label the person has supplied, but to listen, ask, and narrow: to move a diffuse presentation toward an area of clinical concern that warrants closer assessment. This is the work of triage.

Within the product architecture of which this agent is a part, the triage agent occupies Layer 1 of the user journey. Layer 0 is a symptom extractor; Layer 2 is a set of diagnosis-specific agents. Layer 1 receives the user who cannot yet route themselves and, crucially, does not diagnose — it orients. It is a sibling of a clinic-facing screening system rather than a replacement for it: the two are independent, each self-contained, and each evolves on its own. The clinic-facing sibling augments self-report with electrophysiological measurement; the conversational triage agent works from self-report alone, at home, the way a clinician routinely orients a presentation with the PHQ-9, the PCL-5, and an interview before any instrumentation is involved.

1.2 Two questions, asked at 2 a.m.

This paper began with two questions. First: is there an algorithm that can make a genuinely intelligent triage agent through a comfortable conversation? Second: is there a literature that supports it, and is there an algorithm we can contribute that is genuinely new? The answer to both is yes, but the second answer requires care about what “new” means. The components of an intelligent triage agent — adaptive item selection, transdiagnostic dimensional structure, Bayesian sequential reasoning, structured-interview gating, longitudinal monitoring — are individually well established. What is thin in the literature, and what we contribute, is their integration in the medium of comfortable dialogue. The contribution is not a new component. It is an orchestration.

1.3 The gap, confirmed by review

The most recent systematic mapping of the field reviewed 160 studies of AI mental health chatbots published between 2020 and 2024 and classified their architectures as rule-based, machine-learning-based, or LLM-based, proposing a three-tier evaluation framework of bench testing, pilot feasibility, and clinical efficacy. Its central finding is a fragmented landscape with limited rigorous clinical validation, particularly for generative systems: most LLM-based interventions remain in early development phases, and the review urges that clinicians and policy makers distinguish marketing claims from technical realities. A companion scoping review of generative LLMs for mental health found only 16 of 726 screened articles met inclusion criteria, with applications showing initial promise but uncertain effectiveness. The empty ground is therefore not a matter of opinion; it is documented. CDDL is designed to occupy precisely that ground — a conversational triage algorithm whose every move is, by construction, a measurable prediction.

1.4 Contributions of this paper

1. A single verifiable objective for conversational triage: at every turn, select the question that maximizes expected information gain over a live hypothesis board, with a stopping rule tied to the information-gain floor rather than to a fixed turn count.
2. A four-term decomposition of that objective — DiagnosticSeparation, TemporalValue, DimensionalCoverage, and ConversationalCost — with the mathematical form of each term and an account of which terms depend on which parts of a curated knowledge vault.
3. TemporalValue: the functional mechanism by which within-user multi-session memory feeds question selection, so that the same dimension is queried differently depending on whether its signal is novel or recurrent.
4. A constitution / phenotype separation that places the safety floor (red-flag scanning, DSM gate questions, the not-a-substitute boundary) outside the weighted optimization entirely, so that no amount of weight tuning can move it.
5. Three emergent properties of the integrated loop — temporal-aware question selection, confidence-modulated conversation, and information-gain stopping — none of which any single component exhibits alone.
6. A validation protocol: because expected information gain is a number, and the realized reduction in board entropy after a question is also a number, every question the agent asks is a falsifiable prediction.

Throughout, we use the term client rather than patient in all user-facing description, consistent with a non-physician practitioner scope and with a positioning oriented toward orientation and self-understanding rather than diagnosis; the term patient is retained only in the DSM diagnostic context and in legacy data-schema names.

2. Background and Related Work

CDDL stands on four areas of solid ground and one area of thin ground. The thin ground is the contribution surface. We review each in turn and state precisely what CDDL takes from it.

2.1 Adaptive measurement — solid ground for question selection

Operating validated scales by selecting each next item to maximize information is an established field in mental health measurement. The Computerized Adaptive Test for Mental Health (CAT-MH) suite, built on multidimensional item response theory with random-forest diagnostic classification, has been validated against lengthy structured clinical interviews with reliability exceeding that of fixed-length tests, and in primary care it showed diagnostic accuracy comparable to the PHQ-9 and GAD-7 while being delivered in a form patients preferred. The CAT-MH family now spans depression, anxiety, mania, substance use, suicidality, post-traumatic stress, and youth versions. Adaptive measurement is, in effect, the academic twin of conversational triage; the decisive difference is the medium. CAT-MH is form-based, presenting discrete items rated on fixed scales. CDDL is dialogue-based.

The underlying principle is shared: an initial response yields a provisional estimate of a latent trait, that estimate selects the next item, and the procedure continues until the estimation uncertainty falls below a predefined threshold. This is exactly the stopping logic of the CDDL hypothesis board. The principle also exposes the limitation of fixed questionnaires, which assume that every item is equally discriminating and contributes equally — an assumption that is often inaccurate, so that fixed instruments trade precision for standardization. CDDL's strategy of "disassembling the questionnaire and wearing it" — preserving the measurement structure while discarding the fixed item wording — follows directly.

2.2 Transdiagnostic dimensional structure — solid ground for the dimension layer

The Hierarchical Taxonomy of Psychopathology (HiTOP) is a data-driven, hierarchical alternative classification that conceptualizes psychopathology as a set of dimensions organized into progressively broader transdiagnostic spectra. Its structure has been supported by meta-analysis: a model closely resembling the HiTOP framework fit the data well, and the placement of DSM diagnoses within transdiagnostic dimensions was largely confirmed. The Brief HiTOP (B-HiTOP) is the direct precedent for CDDL's Layer 0: an efficient screening instrument that lets a clinician identify areas of clinical concern on HiTOP dimensions that cut across traditional diagnostic boundaries, and use the result to target deeper assessment. That is precisely what the CDDL agent does — dimensional screening, then targeted differential probing of active areas.

HiTOP was also designed from the outset as a living model, conceptualized as evolving and requiring continuous revision to keep pace with new findings. CDDL's curated version evolution shares this philosophy: a model that breathes, but is revised through review rather than drift.

2.3 Bayesian sequential differential reasoning — solid ground for the board and selector

A second tradition treats diagnosis as sequential decision-making under uncertainty. A Bayesian formulation of automated differential diagnosis applies Bayesian inference to the disease-inference step and Bayesian experimental design to the inquiry step, an approach that is interpretable, requires no costly training, and

adapts to new variation without additional effort. This is the mathematical foundation of CDDL's hypothesis board and next-question selector; the properties "interpretable / training-free / adaptable" align directly with the self-contained, curation-based design philosophy.

Structured clinical interview systems contribute the gating idea: such systems are characterized by deductive reasoning and automatic generation of diagnoses, following a predetermined decision tree that requests input judgments on criteria presented sequentially. But the same line of work issues an important warning — computerized sequential systems make global thinking and back-and-forth exploration harder, forcing thought into a sequence. That warning is the reason CDDL's hypothesis board exists. The board forces the agent to hold several hypotheses simultaneously, guarding against the trap of pure sequential reasoning: automation bias and errors of omission.

More recent work frames clinical reasoning as interactive question-asking under incomplete information. The MediQ benchmark and question-asking system establishes that directly prompting state-of-the-art LLMs to ask questions actually degrades clinical reasoning, that adapting LLMs to proactive information-seeking is non-trivial, and that an abstention module which estimates model confidence and decides when to ask more questions improves diagnostic accuracy substantially — though performance still lags an upper bound in which full information is given upfront. Two lessons carry directly into CDDL: a naive "just ask questions" LLM is not a triage agent, and an explicit, confidence-aware mechanism for deciding what to ask and when to stop is required. CDDL's four-term objective and information-gain stopping rule are that mechanism, made explicit.

The most recent line of this work goes further and supplies the structural vocabulary CDDL adopts. Knowledge-graph-anchored diagnostic frameworks pair an information-gain questioning strategy with an explicit, state-tracked diagnostic record that updates a physician's belief as evidence emerges — the same pairing as CDDL's next-question selector and live hypothesis board. Benchmarks for goal-directed diagnostic questioning establish that strategic inquiry, not single-shot accuracy, is the property worth measuring, and that an adaptive sequence of focused questions in which each inquiry depends on the prior answer is what separates an efficient agent from an inefficient one. CDDL is, in these terms, a psychiatric-triage instance of an information-guided inquiry loop, with two additions that the general medical work does not carry: a transdiagnostic dimension layer in place of a flat symptom set, and a within-user temporal axis in place of a single-encounter assumption.

2.4 Conversational LLM mental health assessment — the thin ground

This is the newest and emptiest area, and therefore the contribution surface. Beyond the systematic-review evidence already cited, the emergence of consultation-specialized systems names the two failure modes that CDDL is built to address. Without criteria-grounded clinical support, a conversational system is prone to ungrounded clinical assertions when symptoms are atypical or under-specified, and it struggles to mitigate inquiry drift — topic wandering or low-yield questioning — across multi-turn interaction. Ungrounded assertion is what CDDL's gate-and-vault layer prevents; inquiry drift is what the information-gain question selector prevents. The thinness of this ground is not a weakness of the present work; it is the opportunity the present work is written to take.

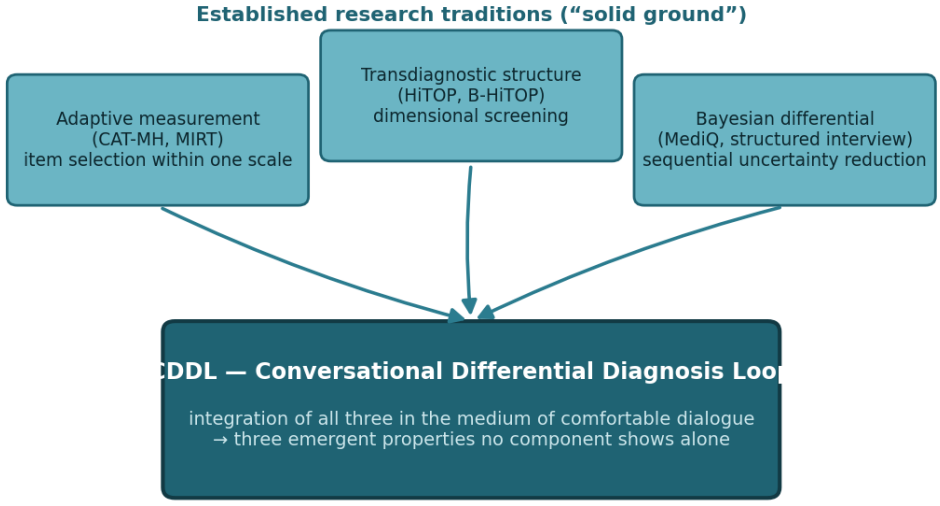
2.5 Longitudinal and measurement-based care — solid ground for the within-user axis

Measurement-based care (MBC) consists of the systematic collection of client data using validated tools and the application of that data to guide and share clinical decision-making, enabling clinicians to monitor symptoms and functioning longitudinally and to support collaborative treatment planning. The case for accumulating signal over multiple sessions, rather than relying on a single snapshot, is reinforced by the limits of recall: retrospective responses are affected by recall failure and response bias, and an end-of-day retrospective report captures only a fraction of the mood variability captured by real-time responses obtained earlier the same day. Longitudinal ecological momentary assessment combined with unsupervised machine learning has been shown to separate groups with high symptom burden from groups with healthy rating patterns and to identify individuals in need of extended care. Together these establish that a single triage snapshot is insufficient and that temporal accumulation separates signal — the academic precedent for an agent that uses time as an ally rather than fighting recall.

Table 1. What CDDL takes from each research tradition.

Tradition	Established result	What CDDL takes	Ground
Adaptive measurement (CAT-MH, MIRT)	Item selection by information; validated against structured interviews; lower burden than fixed tests	The next-question objective and the uncertainty-threshold stopping rule	Solid
Transdiagnostic structure (HiTOP, B-HiTOP)	Dimensions cutting across DSM categories; meta-analytically supported; designed as a living model	The dimension vector of Layer 0; targeted probing of active areas; the living-model philosophy	Solid
Bayesian differential (Bayesian DDx, structured interview)	Inference + experimental design; interpretable, training-free; gating prunes diagnostic bundles	The hypothesis board posterior; gate questions; the board as a guard against pure sequential reasoning	Solid
Interactive question-asking (MediQ)	Naive question-prompting degrades reasoning; an explicit confidence/abstention mechanism is required	The explicit four-term objective; information-gain stopping in place of naive prompting	Solid
Longitudinal / MBC (EMA, measurement-based care)	Single snapshots under-capture; multi-session accumulation separates signal; identifies extended-care need	TemporalValue and the within-user accumulation layer; trait / state / course separation	Solid
Conversational LLM assessment	Documented gap: little clinical validation; ungrounded assertion and inquiry drift are named failure modes	The contribution surface — integration of the above in comfortable dialogue	Thin

Figure 10. Positioning — CDDL as an integration algorithm, not a new component



“The components are established. The integration value is the contribution — and the moat.”

Figure 10. Positioning. The three traditions are established “solid ground.” CDDL is an integration algorithm that combines them in the medium of comfortable dialogue; the integration value, not any single component, is the contribution.

3. The CDDL Architecture

3.1 Seven design principles

The CDDL agent rests on seven agreed principles, stated here as the skeleton the remainder of the paper formalizes.

1. Triage is Layer 1, a separate agent product, a sibling of the clinic-facing screening system — independent, each self-contained, both evolving.
2. The agent is self-contained, operating without external backend infrastructure, embedding its vault within the agent in the manner of a single-file deployment.
3. Evolution is not self-adjustment. The agent proposes pattern candidates; a curator reviews them; a version is incremented. The process is verifiable, reversible, and drift-free.
4. Initial weights come from four sources — DSM-5-TR structure, validated-scale item structure, differential-diagnosis literature, and clinical estimation — each tagged with its source and a confidence level.
5. Constitution and phenotype are separated. The constitution (DSM gate criteria, red-flag logic, safety boundaries) is shared and immutable; the phenotype (dialogue style, question order, low-confidence cell micro-adjustment) is independent and evolvable.
6. Accuracy is built within the agent — within-user multi-session accumulation, depth of differential reasoning, and curated version evolution. Cross-product reference is a v2+ roadmap item, reclassified as an insight tool rather than an accuracy tool.
7. The agent's identity is to orient, not to diagnose. Its output is a prioritized set of multiple recommendations, an explicit statement that it does not replace clinician evaluation, and an immediate branch on any crisis signal.

3.2 The four components of the triage engine

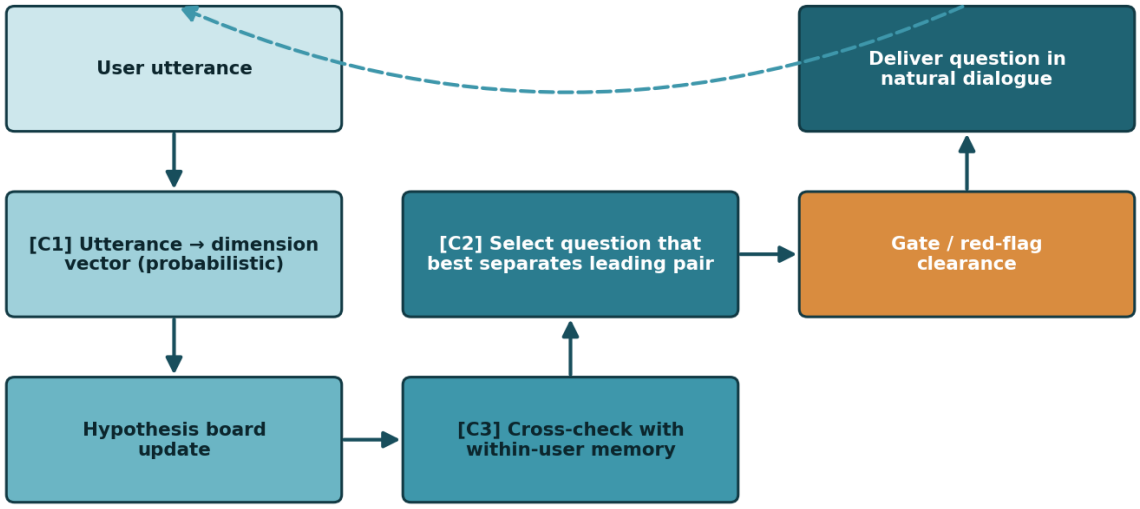
Table 2. The four components of the CDDL triage engine.

Component	Role
Component 1 — Surface-to-Dimension converter	Converts each utterance into a partial update of a transdiagnostic dimension vector. One utterance may contribute to several dimensions simultaneously and with uncertainty.
Component 2 — Hypothesis board	In place of flat scoring, maintains two to four live hypotheses, each carrying {diagnostic area, probability, supporting evidence, decisive unasked question}.
Component 3 — Next-question selector	Each turn, selects the single question that best separates the current hypotheses by maximizing expected information gain. This is the heart of the agent's intelligence.
Component 4 — Gate and safety layer	A red-flag scanner (every turn, top-priority branch) plus time-axis questions (mapping course) — operating in parallel with, and above, the reasoning loop.

Two of these components map onto the agent's discretion — its intelligence — and two onto a fixed floor. The discretionary region covers how deeply to probe an active area, the order of questions, whether to revisit a silent area, and conversational tone. The fixed floor covers the red-flag scan (every turn, unconditionally), the gate questions (which cannot be skipped), and the not-a-substitute-for-clinician-evaluation boundary. This discretion / floor distinction is the seed of the constitution / phenotype separation formalized in Section 6.

Figure 1. The CDDL loop — structure of a single conversational turn

next utterance → loop



Key: Contribution 3 (temporal memory) feeds Contribution 2 (question selection) — the three contributions are not parallel; one is the input of another.

Figure 1. The CDDL loop — the structure of a single conversational turn. The defining feature is that Contribution 3 (temporal memory) feeds Contribution 2 (question selection): the three contributions are not parallel; one is the input of another.

3.3 The loop — one turn

A single CDDL turn proceeds as follows. The user utterance enters Component 1 and is converted into a probabilistic, non-committal dimension vector. The hypothesis board is updated. The updated board is cross-checked against within-user memory — is this a signal heard for the first time, or a recurring one? The next-question selector then chooses, from the board, the question that best separates the most closely competing pair of hypotheses. That candidate passes the gate and red-flag layer. The question is delivered in natural dialogue, and the next utterance re-enters the loop.

> *The defining structural fact: Contribution 3 (temporal memory) is fed into Contribution 2 (question selection). The three contributions are not three things sitting side by side; one is the input of another. That is what "integration" means in this paper, and it is why the loop — not any single component — is the contribution.*

4. The CDDL Function

4.1 The heart of the loop: the next-question selection function

The entire CDDL algorithm reduces to a single line.

```
NextQuestion(state) = argmax_q  ExpectedInfoGain(q, state)
```

Here state is everything accumulated so far — the dimension vector, the hypothesis board, the within-user history, the dialogue state — and q is a candidate question. At every turn, the agent computes, for every candidate question, how much that question is expected to reduce the uncertainty of the hypothesis board, and selects the question that reduces it most.

> *Why this is verifiable: ExpectedInfoGain is a number, and after the question is asked the realized reduction in board uncertainty is also a number. Predicted versus observed can be compared. The algorithm is therefore falsifiable — it is science, not assertion. This is the direct answer to the documented gap that most LLM mental-health systems are unvalidated.*

The objective decomposes into four terms:

```
ExpectedInfoGain(q, state) =
    w₁ · DiagnosticSeparation(q, hypothesis_board)        [Contribution 2]
  + w₂ · TemporalValue(q, user_history)                   [Contribution 3]
  + w₃ · DimensionalCoverage(q, dimension_vector)          [Contribution 1]
  − w₄ · ConversationalCost(q, dialogue_state)            [comfort penalty]
```

Figure 2. Decomposition of the next-question objective into four terms

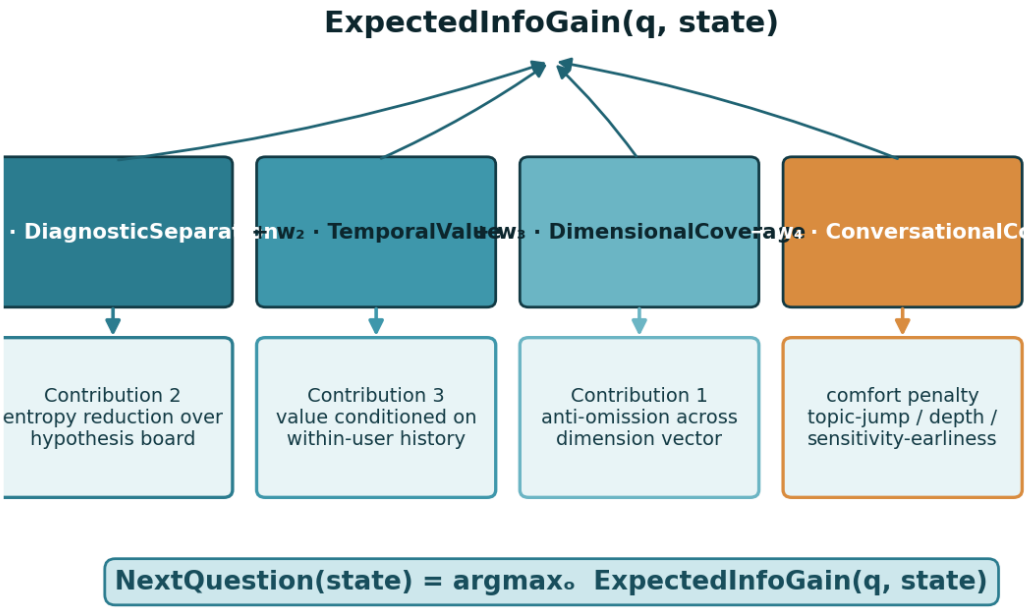


Figure 2. Decomposition of the next-question objective into four terms. Three terms are additive contributions to information gain; ConversationalCost is subtracted as a comfort penalty so the agent does not behave like an interrogation.

4.2 Term 1 — DiagnosticSeparation

This is the most solid term and is almost independent of vault structure. The hypothesis board is a probability distribution P(h) over diagnostic areas h ∈ {MDD, PTSD, GAD, ADHD, …}. Board uncertainty is its entropy:

```
H(board) = − Σₕ P(h) · log P(h)
```

The value of a question q is the expected reduction in entropy it produces:

```
DiagnosticSeparation(q) = H(board) − Σₐ P(a) · H(board | a)
```

Here P(a) is the probability that the user gives answer a, estimated from the current board, and H(board | a) is the entropy of the board after a Bayesian update on answer a. This is the expected information gain of information theory applied directly. The Bayesian update $P(h|a) \propto P(a|h) \cdot P(h)$ draws exactly one quantity from the vault: the likelihood P(a|h), the probability that a person with diagnostic area h gives answer a to question q.

4.2.1 Resolving the likelihood from a descriptive vault

In the current curated vault, the separating-question library stores, for each question, a descriptive mapping — “this answer is suggestive of PTSD” — rather than a numeric likelihood. A natural assumption is that an LLM can simply convert the description to a number on the fly. It cannot, and the distinction matters. Translation is

automatic: an LLM can read “suggestive of PTSD” and infer the direction of evidence. Calibration is not automatic: P(a|h) is a magnitude, not a direction — 0.7 versus 0.4 — and a magnitude produced by an LLM on the fly is a hallucinated number, carrying neither a source nor a confidence tag, in direct violation of the four-source, source-and-confidence-tagged weighting principle.

The resolution is therefore a v1 prerequisite, but a bounded one. The descriptive mappings are curated into a three-level ordinal likelihood — weak ≈ 0.2, moderate ≈ 0.5, strong ≈ 0.8 — each carrying its source tag. A Bayesian update is order-stable: it does not require a precise 0.73, only a correct ordering. This curation is absorbed into the same workflow that builds the diagnosis-specific vaults, not a separate research effort.

> *Validation of Term 1: after a question is asked, measure the realized H(board) and compare it with the predicted reduction. A question predicted to cut entropy by 0.4 bits that realizes 0.05 has a mis-estimated separating power — it becomes a curation candidate.*

4.3 Term 2 — TemporalValue

TemporalValue is the newest part of the function and the functional substance of the claim that Contribution 3 feeds Contribution 2. Its core idea: the same question q has a different value depending on the temporal state, in the user’s history, of the dimension d that q addresses. The temporal state of dimension d is summarized by three numbers:

- seen(d) — the number of times the signal for d has appeared in past sessions;
- variance(d) — the between-session variance of those signals;
- trend(d) — the temporal slope of those signals (large if monotonically increasing or decreasing, near zero if erratic).

Questions are typed: q_state (“how is it now?”), q_course (“have you always been this way, or is it recent?”), and q_onset (“when did it start?”). TemporalValue is then a match on question type and temporal state:

```
TemporalValue(q, history) = match q.type:
  q_state   →  high  if seen(d) = 0          (novel signal → establish current
state)
             low   if seen(d) large       (already seen repeatedly → asking “now?”
wastes a turn)
  q_course  →  high  if seen(d) ≥ 2 and variance(d) low   (recurrent + stable →
trait candidate; course is decisive)
             high  if seen(d) ≥ 2 and trend(d) large    (recurrent + trending →
course is itself the axis)
             low   if seen(d) = 0                       (asking course of a first-
time signal → user is disoriented)
  q_onset   →  high  if seen(d) ≥ 1 and onset undetermined
             low   otherwise
```

This is the functional substance of “Contribution 3 feeds Contribution 2.” TemporalValue takes both the question type and the dimension’s temporal state as input, so that the same dimension scores toward q_state

when it is novel and toward q_course when it is recurrent. The question selector (Term 1) chooses what to ask; TemporalValue chooses from which temporal angle to ask it.

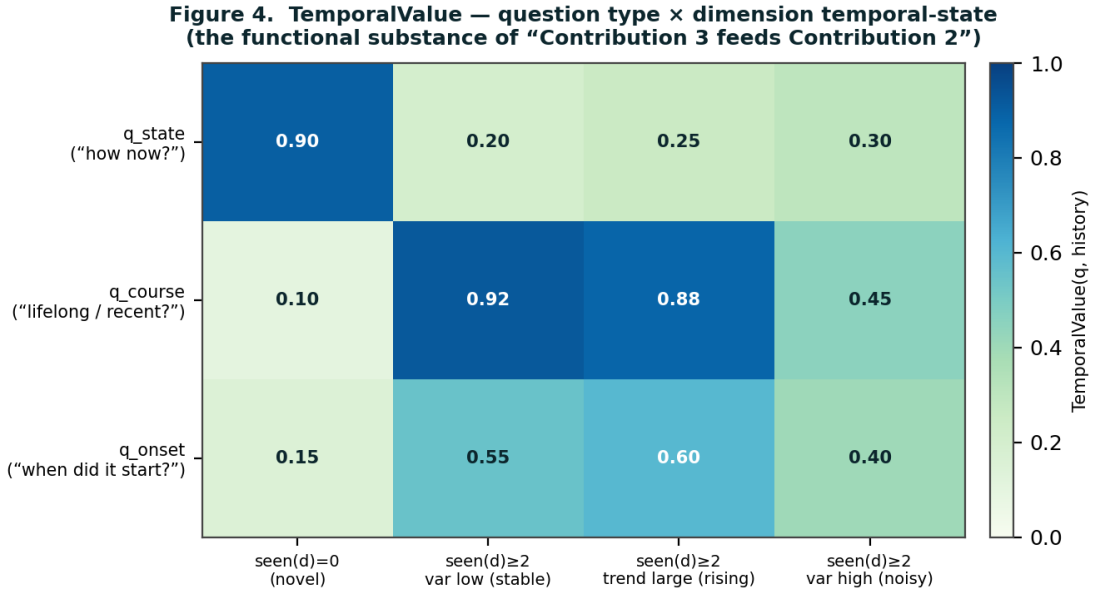


Figure 4. TemporalValue as a decision matrix — question type crossed with the temporal state of the dimension. A novel signal scores toward q_state; a recurrent, stable signal scores toward q_course. This is the functional mechanism by which temporal memory reshapes question selection.

4.3.1 The within-user store and trait / state / course

Computing variance(d) and trend(d) requires the within-user store to hold a per-dimension session time series. The store therefore retains both the full per-dimension scores of each session and the final diagnostic estimate. With the time series present, TemporalValue runs in its complete form; were only the final estimate retained, it would fall back to a reduced form using seen(d) alone — functional, but with weakened trait / course discrimination. The complete form is the one specified here.

The accumulation has a further purpose. Across sessions, signal patterns sort into three classes: stable repetition indicates a trait; erratic fluctuation indicates a state or noise; change over time indicates a course. The course axis is central to diagnostic differentiation — it separates episodic (MDD) from chronic (persistent depressive disorder) from lifelong (ADHD) from post-event (adjustment disorder) presentations — and it is an axis that dimension scores alone can never capture. The agent does not attempt to decide everything in a single session; it accumulates judgment to raise accuracy. That is what a one-shot instrument cannot do and an agent can.

Figure 9. Within-user multi-session time series — the same accumulation separates trait, state and course

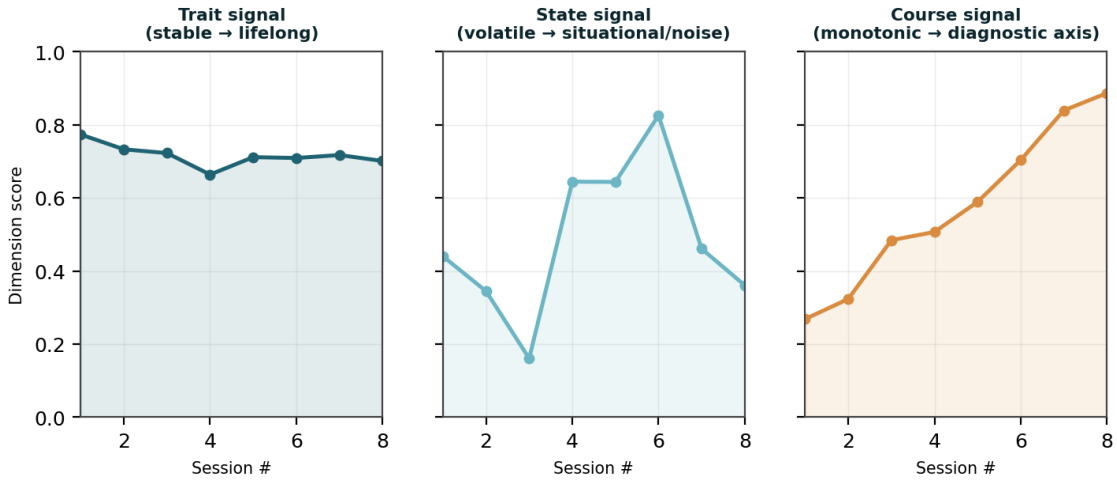


Figure 9. Within-user multi-session time series. The same accumulation separates a stable trait signal, a volatile state signal, and a monotonic course signal — and the course signal is a diagnostic axis unavailable to any single-session score.

> *Validation of Term 2: after a q_course question is asked, check whether the dimension settles into a stable trait / state classification in the following session. If it does, the TemporalValue estimate was correct.*

4.4 Term 3 — DimensionalCoverage

DimensionalCoverage prevents omission. It rewards questions that touch dimensions not yet covered:

```
DimensionalCoverage(q, dimension_vector) = Σ_{d ∈ dims touched by q}  untouched(d)
· gate_weight(d)
```

- untouched(d) = 1 if dimension d still has no signal, else 0;
- gate_weight(d) is large if d is a DSM gate dimension — trauma exposure, manic episode, psychosis — because omitting it leaves an entire diagnostic bundle unclosed.

The effect: when the agent would otherwise probe only an active area and neglect a silent one, a bonus on untouched-dimension questions pulls it back. This term has almost no vault dependency — it needs only the dimension list and the gate-dimension flags, both of which come from the DSM. Its validation is a coverage check: at session end, is every dimension classified as active, excluded, or at-risk? If even one untouched dimension remains, that session has failed coverage.

4.5 Term 4 — ConversationalCost

ConversationalCost is the subtracted term — the place where “comfortable conversation” is written into the function. If the agent chases information gain alone, it becomes an interrogation; ConversationalCost subtracts to prevent that.

```
ConversationalCost(q, dialogue_state) =
```

```
    c₁ · topic_jump(q, last_q)              (distance from the previous topic —
abrupt-switch penalty)
  + c₂ · depth_overrun(q, current_topic)    (staying on one topic too long — over-
probe penalty)
  + c₃ · sensitivity(q) · earliness         (a sensitive question early in the
dialogue — penalty decays later)
```

- sensitivity(q) is the per-question sensitivity tag — high for trauma and self-harm questions;
- earliness = 1 / current_turn_number, large at the start of the conversation.

4.5.1 The sensitivity tag and the constitutional exception

The vault does not currently carry a per-question sensitivity tag, so tagging is a v1 prerequisite: an AI model produces a first-pass tag, and a curator reviews it. There is, however, a safety caveat. The sensitivity tag of a red-flag, self-harm, or trauma question is not left to AI estimation — it is fixed on the constitutional side. The AI proposes "this question is of medium sensitivity"; it does not classify "this is a red-flag question." Human judgment fixes the latter. This prevents the failure mode in which ConversationalCost penalizes, and therefore suppresses, a red-flag question.

> *Crucially, ConversationalCost is not constitutional. A red-flag question, however high its sensitivity, is asked unconditionally — ConversationalCost is ignored for it. The penalty shapes a comfortable conversation; it never shapes the safety floor.*

Table 3. The four terms — form, vault dependency, and validation signal.

Term	Direction	Vault dependency	Validation signal
DiagnosticSeparation	+ (information gain)	Likelihood P(a|h); descriptive → three-level ordinal (v1 prerequisite, bounded)	Predicted vs. realized H(board) reduction
TemporalValue	+ (information gain)	Within-user per-dimension time series (present → complete form)	Does the dimension settle into stable trait / state next session?
DimensionalCoverage	+ (information gain)	Almost none — dimension list and gate flags from DSM	At session end, is every dimension classified?
ConversationalCost	− (comfort penalty)	Per-question sensitivity tag (AI first pass + curator; red-flag fixed by human)	User attrition; per-turn response length trend

5. Weights, the Per-Turn Computation, and a Worked Trace

5.1 The weights w_1 w_2 w_3 w_4 — handling "it differs by person"

The weights are not fixed constants. The objection that the right balance differs by person is correct, and it is handled in two ways rather than by pretending a single constant exists.

1. The constitution sits outside the weights. Red-flag detection and gate questions do not enter the weighted sum at all; they are unconditional branches above the argmax. However the weights are tuned, the safety line does not move.
2. The weights are curated parameters. Their initial values are source-and-confidence-tagged estimates. As data accumulates, the agent proposes — "triage was more accurate under this weight combination" — and a curator reviews and increments the version. This is not self-adjustment. The change in weights is itself a validation target: whether the v2 weights genuinely outperformed the v1 weights is scored on held-out data.

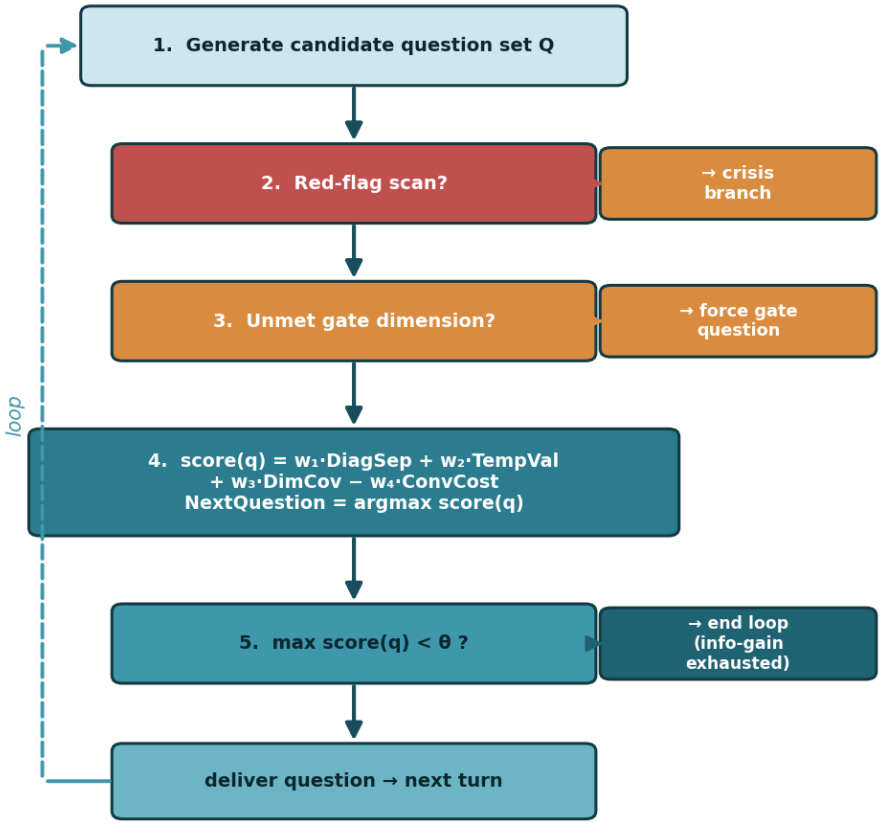
5.2 The per-turn computation

Assembling the terms, a full CDDL turn is computed as follows.

```
Each turn:
  1. Generate the candidate question set Q
     (separating-question library + gate questions + time-axis questions)
  2. Red-flag scan — if triggered, skip the argmax and branch to crisis
[CONSTITUTION]
  3. If any gate dimension is unmet → force the corresponding gate question
[CONSTITUTION]
  4. Otherwise, for each q ∈ Q:
       score(q) = w₁·DiagnosticSeparation(q) + w₂·TemporalValue(q)
                + w₃·DimensionalCoverage(q) − w₄·ConversationalCost(q)
     NextQuestion = argmax score(q)
  5. Stopping condition: if max score(q) < θ, end the loop (information gain
exhausted)
```

Steps 2 and 3 are the constitution — outside the weights, unconditional. Step 4 is the curated part. Step 5 makes dialogue length self-adjust to case complexity through the single threshold θ: a simple case terminates in roughly eight turns, a complex comorbid case in roughly eighteen.

Figure 12. Full per-turn computation of the CDDL loop



Steps 2–3 = constitution (outside weights, unconditional). Step 4 = curated. Step 5 = self-adjusting length.

Figure 12. The full per-turn computation. Steps 2 and 3 are constitutional — outside the weighted sum and unconditional. Step 4 is the curated weighted objective. Step 5 is the self-adjusting stopping rule.

5.3 A worked dialogue trace

The following condensed trace illustrates one path through the loop. It is illustrative, not a transcript of a validated interaction; it shows how the four terms interact turn by turn.

Table 4. A worked CDDL trace — an MDD–PTSD differential (condensed).

Turn	User utterance (paraphrased)	Board after update	Term that dominated q-selection	Question delivered
1	"Nothing feels enjoyable lately and I'm exhausted."	MDD .34 / PTSD .26 / GAD .22 / ADHD .18	DimensionalCoverage — most dimensions still untouched	"When you say nothing feels enjoyable — is that everything, or particular things?"
3	"It's mostly when something reminds me of last year."	MDD .40 / PTSD .30 / GAD .15 / ADHD .14	DiagnosticSeparation — anhedonia vs. avoidance-linked numbing is the live split	"Is the numbness there all the time, or does it spike when you avoid certain memories?"
6	"It spikes. And honestly I've felt low for years, not just now."	MDD .45 / PTSD .36 / GAD .12 / ADHD .08	TemporalValue — a recurrent low-mood signal; course is now decisive	"When you say years — has it been fairly constant, or does it come and go in episodes?"
8	"Comes and goes, but there's always some baseline heaviness."	MDD .44 / PTSD .40 / GAD .08 / ADHD .08	Gate (constitution) — trauma-exposure gate not yet cleared	"Earlier you mentioned last year — was there a specific event

Turn	User utterance (paraphrased)	Board after update	Term that dominated q-selection	Question delivered
				that still feels unfinished?"
11	(content indicating trauma exposure; no red-flag)	MDD .43 / PTSD .43 — comorbid profile, board stays live	Information-gain stopping — max score(q) approaching θ	— (loop nears termination; prioritized orientation prepared)

The trace makes the integration visible. No single term drives the conversation: DimensionalCoverage opens it, DiagnosticSeparation sharpens the live split, TemporalValue redirects toward course once a signal recurs, the gate layer asserts itself unconditionally, and information-gain stopping ends the loop when further questions would yield little. The board does not collapse to one label; it resolves to a prioritized, comorbid orientation — which is the honest output of triage.

Figure 5. Live hypothesis board — posteriors evolving toward an MDD–PTSD comorbid profile rather than collapsing to a single category

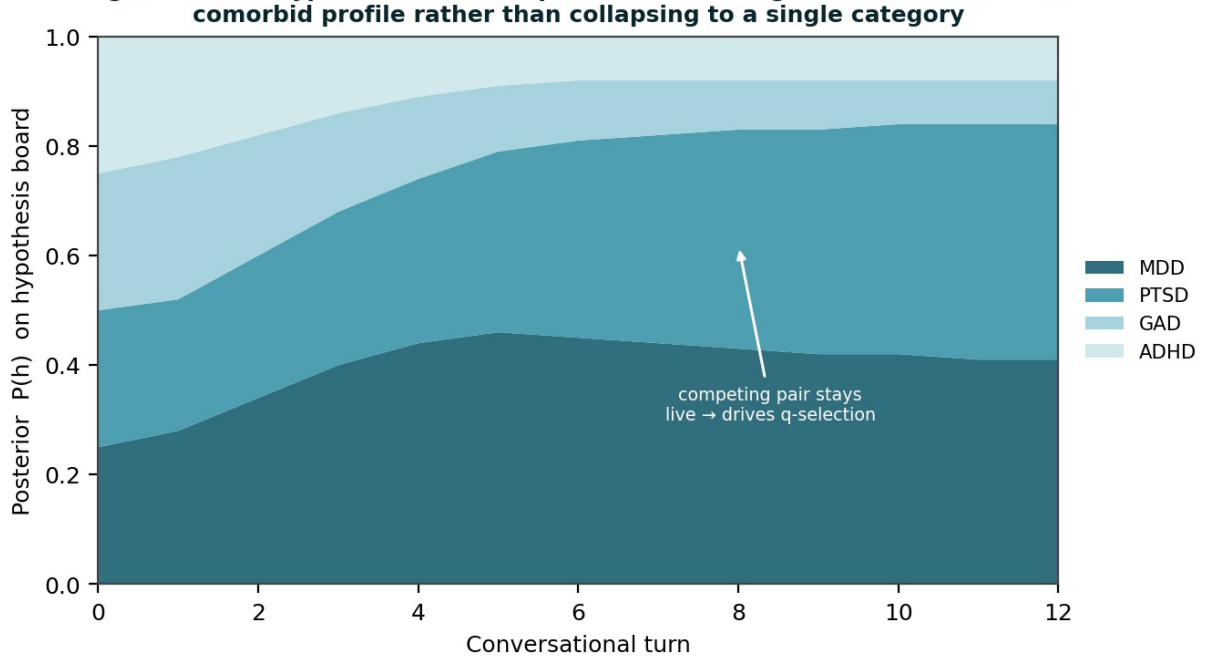


Figure 5. The live hypothesis board across the trace. Posteriors evolve toward an MDD–PTSD comorbid profile rather than collapsing to a single category; the competing pair stays live and drives question selection.

6. Constitution, Phenotype, and Curated Evolution

6.1 Why self-adjusting learning is excluded

Self-adjustment — the agent changing its own weights from user data — is prohibited in the diagnostic domain, for three reasons. Drift: a biased user population tilts the triage criteria themselves. Absence of ground truth: self-report estimates pull on one another and can become wrong together. Accountability vacuum: being responsible for something one does not control. Medical diagnostic AI is governed by a locked-algorithm principle, and CDDL adheres to it.

6.2 Curated version evolution

In place of self-adjustment, the agent improves through a curated pipeline. Real data accumulates. The agent discovers and proposes weak-point patterns — not weight changes, but observations: “this separating question is weak,” “this utterance should also have updated this dimension.” A curator reviews these proposals as one would refine a diagnosis-specific vault. Only the phenotype is revised; the constitution — DSM gates, red-flag logic — is not changed by data. A version is incremented, verifiable and reversible.

This is how a genuine expert improves with experience. A good clinician does not unilaterally change their criteria after a hundred cases; they notice a pattern, subject it to review, and update their clinical sense in a validated form. HiTOP itself is a model that breathes but is revised through review — the same principle.

Figure 8. The curation pipeline — how the bot gets smarter safely

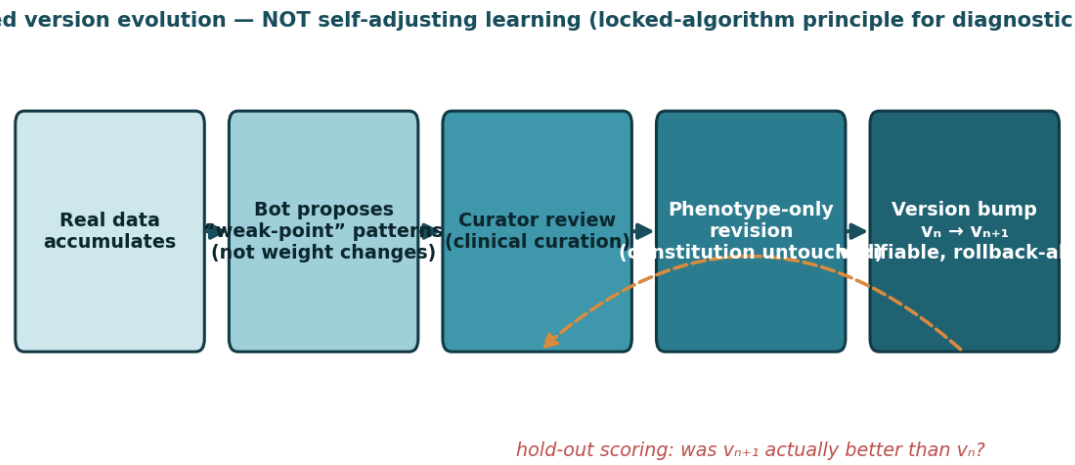


Figure 8. The curation pipeline. The agent proposes weak-point patterns; a curator reviews; only the phenotype is revised; a version is incremented. A hold-out scoring loop checks whether each new version was genuinely better than the last.

6.3 The constitution / phenotype separation

The constitution is shared and immutable: the red-flag scanner that runs every turn as the top-priority branch, the DSM gate questions that cannot be skipped, and the explicit not-a-substitute-for-clinician-evaluation boundary. It sits outside the weighted sum entirely. The phenotype is independent per agent and curation-evolvable: dialogue style and question order, the curated weights w_1–w_4, separating-question library tuning,

and low-confidence cell micro-adjustment. The argmax operates only on the phenotype; the constitution sits above it.

> *The single most important safety property of CDDL: the safety floor is not a tunable parameter. No combination of weights, no curation increment, and no accumulated data can move the red-flag scan, the gate questions, or the not-a-substitute boundary. They are structurally outside the optimization.*

Figure 7. Constitution / phenotype separation — the safety floor is not a tunable parameter

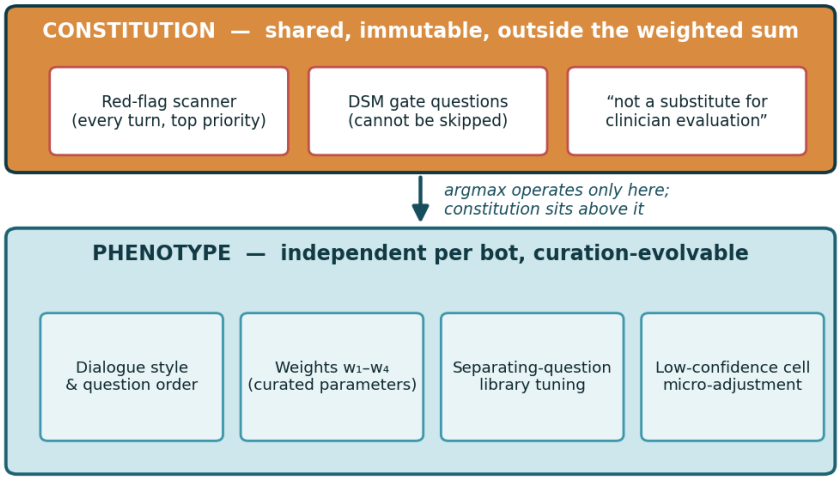


Figure 7. Constitution / phenotype separation. The constitution — red-flag scanning, DSM gate questions, the not-a-substitute boundary — is shared, immutable, and outside the weighted sum. The phenotype — dialogue style, curated weights, library tuning — is independent per agent and curation-evolvable.

7. Three Emergent Properties of the Integrated Loop

The claim of this paper is not that any component is new. It is that integrating the components into a single conversational loop produces properties that no component exhibits alone. We identify three.

7.1 Temporal-aware question selection

The same dimension is queried differently depending on its temporal state: q_state if the signal is novel, q_course if it is recurrent. The question selector consults temporal memory and changes the type of question itself — not merely its priority. Neither CAT-MH nor a naive question-asking LLM does this; CAT-MH selects items within a scale, and a naive LLM has no structured temporal memory to consult. Term 2 of the objective is the formal expression of this property.

7.2 Confidence-modulated conversation

The confidence of the hypothesis board changes how the agent speaks. When the board is near-tied — say MDD .52 / PTSD .48 — the agent enters an exploratory mode, probing more gently and broadly. When the board has hardened, it enters a confirmatory mode. The agent's mode detector links its reasoning state to its conversational register, so that the dialogue style is a readout of diagnostic confidence rather than a fixed persona.

7.2.1 Formalizing confidence modulation in the function

This property can be made explicit in the objective rather than left as an emergent side effect. We define a board-confidence scalar from the same entropy already computed for Term 1:

```
confidence(board) = 1 − H(board) / H_max,    H_max = log(number of live
hypotheses)
```

confidence is 0 when the board is maximally uncertain (all hypotheses equiprobable) and approaches 1 as the board hardens. It modulates the ConversationalCost term: when confidence is low, the agent should tolerate gentler, more exploratory questions and penalize abrupt, high-sensitivity probing more heavily; when confidence is high, the agent can afford more direct confirmatory questions. Concretely, the earliness-style decay in ConversationalCost is generalized so that the sensitivity penalty scales with (1 − confidence(board)):

```
ConversationalCost′(q) = c₁·topic_jump(q) + c₂·depth_overrun(q)
                 + c₃·sensitivity(q)·earliness·(1 − confidence(board))
```

The effect is that an unresolved board (low confidence) makes the agent conversationally cautious — it explores before it presses — while a resolved board (high confidence) lifts that caution and lets the agent confirm. The mode detector and the function now agree: confidence-modulated conversation is no longer only an emergent behavior but a written term, and it remains within the phenotype — the constitutional red-flag branch still ignores ConversationalCost′ entirely.

7.3 Information-gain stopping

The agent does not fill a fixed number of turns. It stops when the information gain of the next question falls below the threshold θ. Dialogue length self-adjusts to case complexity — a simple single-domain presentation closes in roughly eight turns, a complex comorbid presentation in roughly eighteen. Step 5 of the per-turn computation is this property. It is also the direct structural answer to inquiry drift: a low-yield question, by definition, has low information gain and is not selected; when no question clears θ, the loop ends rather than wandering.

Figure 6. Information-gain stopping — dialogue length self-adjusts to case complexity (no fixed turn count)

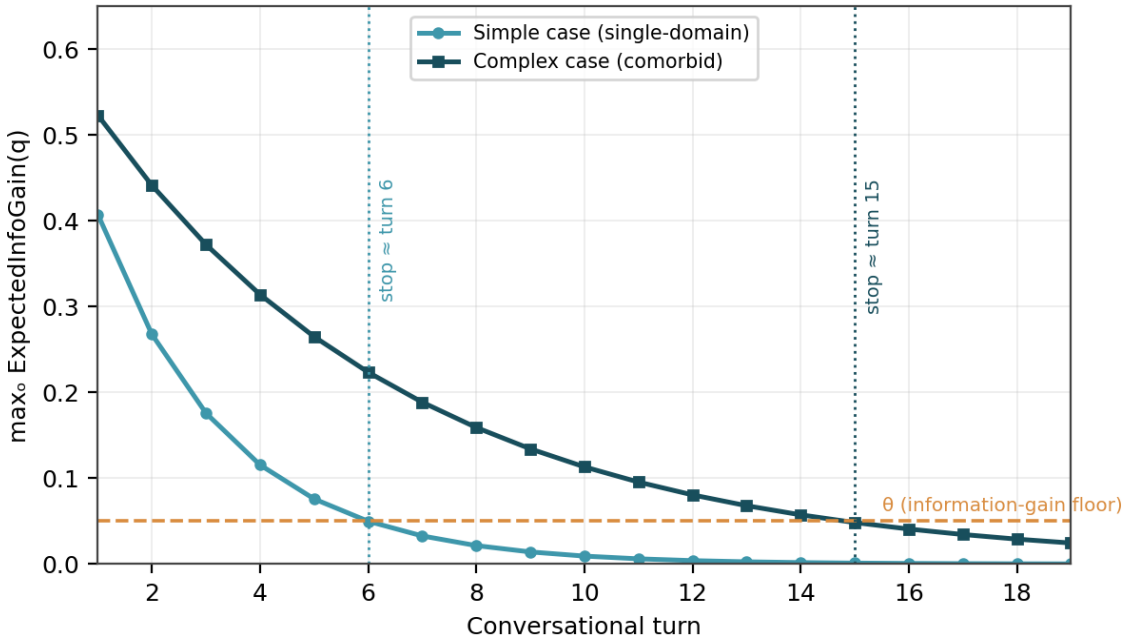
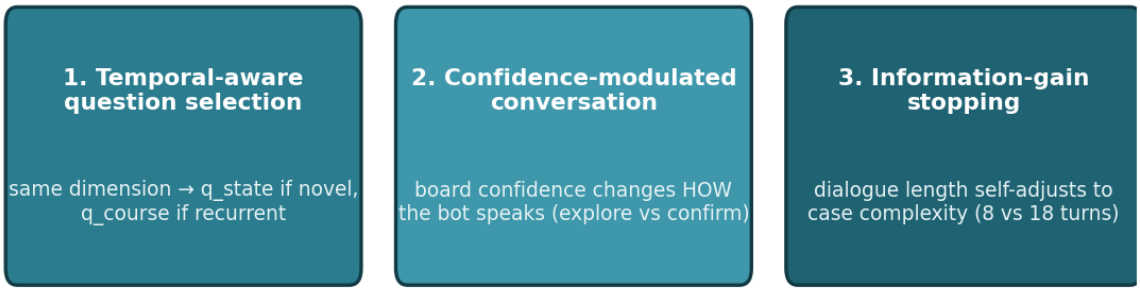


Figure 6. Information-gain stopping. The maximum expected information gain decays as the board resolves; the loop ends when it falls below θ. A simple case crosses θ early; a complex comorbid case sustains gain longer. Dialogue length is an output, not a setting.

Figure 11. Three emergent properties of the integrated loop



Emergent under CDDL integration only — no individual component (CAT-MH, HiTOP, MediQ) exhibits these

Figure 11. The three emergent properties of the integrated loop — temporal-aware question selection, confidence-modulated conversation, and information-gain stopping — none of which any single component (CAT-MH, HiTOP, MediQ) exhibits alone.

The one-line statement of the paper: the individual measurement techniques — IRT adaptive measurement, transdiagnostic dimensions, Bayesian differential reasoning — are established. What we show is that when they are integrated into a single conversational loop, three emergent properties appear that no component exhibits alone. This is the position from which the documented gap — that only a small fraction of LLM mental-health chatbots have been clinically validated — can be addressed directly.

8. Validation Protocol

Because every term of the objective is a number, CDDL is validatable term by term and as a whole. We state the protocol without claiming results: this paper specifies an algorithm and its falsification conditions; an empirical study is the next step.

8.1 Term-level validation

Table 5. Term-level falsification conditions.

Term	Prediction	Measurement	Falsified if
DiagnosticSeparation	Asking q reduces H(board) by the predicted amount	Realized H(board) after the Bayesian update	Realized reduction systematically departs from predicted (e.g., predicted 0.4 bits, realized 0.05)
TemporalValue	A q_course question on a recurrent signal yields a stable trait / state classification	Trait / state classification stability in the following session	q_course questions do not improve next-session classification stability over q_state
DimensionalCoverage	Every dimension is classified by session end	Count of untouched dimensions at termination	Untouched dimensions persist at termination at a non-trivial rate
ConversationalCost	Subtracting the comfort penalty reduces attrition and sustains response length	User attrition rate; per-turn response-length trend	Attrition and response-length decay are unchanged with the penalty removed

8.2 Loop-level validation

- Convergence: does H(board) decline monotonically faster under the CDDL objective than under a random-order or fixed-questionnaire baseline (Figure 3)?
- Stopping calibration: does the realized turn count scale with independently-rated case complexity, and does θ separate simple from comorbid cases (Figure 6)?
- Curation gain: scored on held-out data, does version v_{n+1} genuinely outperform v_n (Figure 8)? A version increment that does not is rolled back.
- Concordance: does the prioritized orientation produced by CDDL agree with an independent clinician's area-of-concern judgment — not as a diagnosis, but as a routing decision?
- Safety: the constitutional branches (red-flag, gate) are audited separately and unconditionally; they are not scored against the weighted objective because they are not part of it.

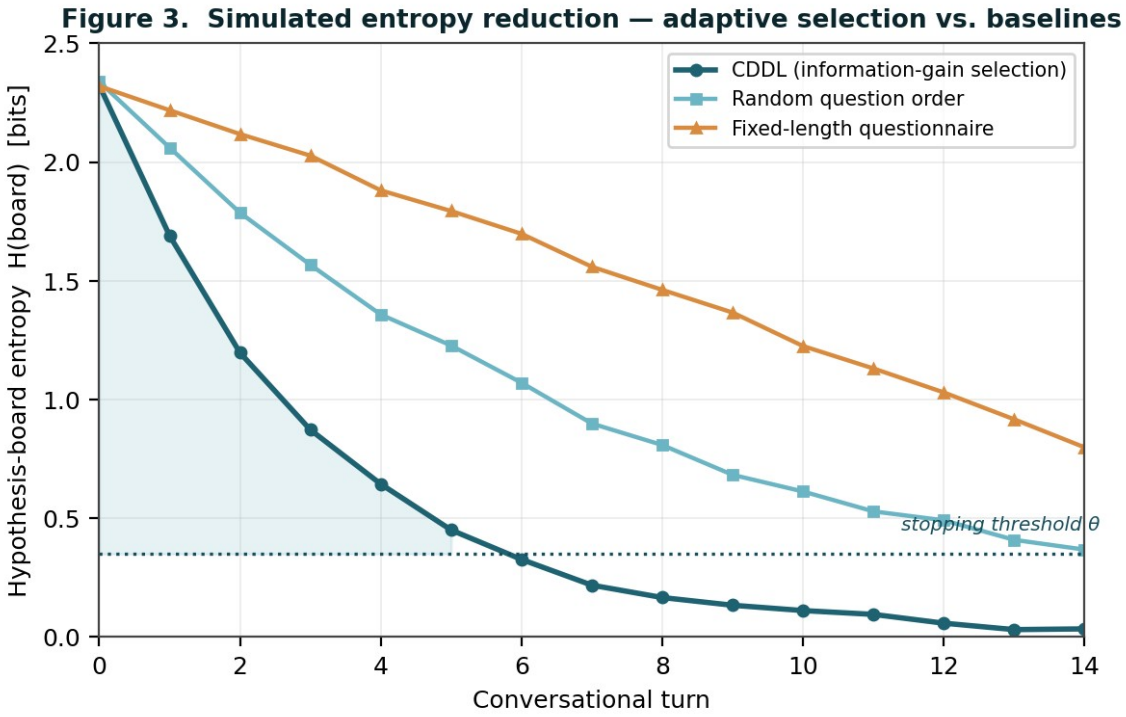


Figure 3. Simulated entropy reduction. Under information-gain selection, board entropy falls toward the stopping threshold faster than under random question order or a fixed-length questionnaire. The figure is a simulation illustrating the predicted dynamics, not empirical data; the validation protocol of Section 8 specifies how the prediction would be tested.

The three baselines in Figure 3 are chosen deliberately. The fixed-length questionnaire is the incumbent: it asks the same items in the same order regardless of what the person has already said, and its entropy curve is the slowest because it spends turns on items that, for this particular board, are non-discriminating. The random-order baseline isolates the contribution of selection from the contribution of simply asking questions at all — if CDDL did not outperform random ordering, the four-term objective would be adding nothing. The CDDL curve is predicted to fall fastest and to flatten as it approaches the stopping threshold, because once the board has resolved, no remaining question clears θ. The gap between the curves is the quantity the empirical study measures; the protocol does not assume the gap, it tests for it, and a CDDL curve that did not separate from random ordering would falsify the central claim.

Loop-level validation also has an order. Term-level falsification comes first, because a loop assembled from mis-estimated terms cannot be diagnosed at the loop level — one would not know which term failed. Convergence and stopping calibration come next, since they test the loop as a whole against the baselines. Curation gain is tested last and continuously, because it is not a one-time result but a standing property: every version increment is a new prediction that must outperform its predecessor on held-out data or be rolled back. This ordering is itself part of the protocol.

8.3 What this paper does and does not claim

This paper specifies an integration algorithm and the conditions under which it would be falsified. It does not report a clinical trial, and it does not claim diagnostic accuracy. The agent it describes orients; it does not diagnose. Its output is a prioritized set of areas of clinical concern, an explicit statement that it does not replace

clinician evaluation, and an immediate crisis branch — and the premise that self-report alone supports orientation at this level is the same premise under which clinicians orient presentations daily with the PHQ-9, the PCL-5, and an interview.

9. Scope, Limitations, and Future Work

9.1 What v1 excludes

The energy of an early idea pushes toward connecting everything. To make v1 shippable, three things are deliberately excluded.

- Cross-product reference — mutual reference between the triage agent and the clinic-facing screening sibling — is deferred to v2+. It is reclassified from an accuracy tool to an insight tool: a discrepancy between the two products (a formal context versus a private one) is itself a clinical signal, such as an avoidance pattern, but realizing it requires account linkage, shared storage, and a privacy structure that come later.
- Self-adjusting learning is permanently excluded, replaced by curated version evolution (Section 6).
- Electrophysiology-based triage is not part of this agent. Its users do not run QEEG at home. This is not a limitation but the natural point of differentiation from the clinic-facing sibling: the agent orients from self-report at home; the sibling adds electrophysiology in a clinical setting.

9.2 Limitations

- The likelihoods P(a|h) begin as a three-level ordinal curation, not as empirically calibrated values; empirical recalibration is itself a validation target.
- The simulation in Figure 3 illustrates predicted dynamics; it is not evidence of clinical performance.
- The worked trace in Table 4 is illustrative; it is not drawn from a validated interaction corpus.
- Within-user accumulation assumes returning users; a single-session user receives the reduced form of TemporalValue.
- The agent depends on a curator. Curation is a strength for safety and a constraint on scaling; the pipeline of Section 6 is designed to keep the curation load light, but it does not eliminate it.

9.3 Future work

1. An empirical study executing the validation protocol of Section 8, beginning with term-level falsification and loop-level convergence.
2. Specifying the constitution / phenotype boundary per DSM diagnosis — an explicit list of what is shared-immutable and what is curation-evolvable.
3. The v1 minimal-viable specification — the smallest implementation of Components 1–4 that ships.
4. The curation pipeline tooling — how the agent raises a proposal and how a curator reviews it with minimal load.
5. Building the vaults for the three contributions, applying the same method already used to build the diagnosis-specific vaults.

10. Conclusion

A person who is psychologically unwell often cannot say where it hurts. The work of triage is to listen, ask, and narrow — and to do so, in a conversational agent, without diagnosing and without interrogating. This paper has specified the Conversational Differential Diagnosis Loop, an algorithm for exactly that task.

The literature is clear on two points. The components of an intelligent triage agent are established: adaptive measurement, transdiagnostic dimensional structure, Bayesian sequential differential reasoning, structured-interview gating, and longitudinal measurement-based care. And the integration of those components in a comfortable conversation is thin — a systematic review of 160 studies documents that only a small fraction of LLM-based mental health systems have undergone clinical efficacy testing. CDDL is written to occupy that documented gap.

Its contribution is not a new component but an orchestration: a single verifiable objective — select the question that maximizes expected information gain over a live hypothesis board — decomposed into four terms whose forms, vault dependencies, and falsification conditions are all specified; a constitution / phenotype separation that places the safety floor structurally outside the optimization; a curated evolution pipeline that lets the agent improve without drift; and three emergent properties — temporal-aware question selection, confidence-modulated conversation, and information-gain stopping — that no component exhibits alone.

The components can be copied. The integration value cannot — because the loop has curation, clinical judgment, and a curated vault embedded inside it. The components are established; the integration is the contribution; and, for a conversational triage agent, the integration is the defensible asset.

References

Adaptive Testing Technologies. (2024). CAT-MH & K-CAT: Validated computerized adaptive tests for mental health. Retrieved from https://adaptivetestingtechnologies.com/

Beiser, D. G., Vu, M., & Gibbons, R. D. (2019). Validation of the Computerized Adaptive Test for Mental Health in primary care. Annals of Family Medicine, 17(1), 23–30. https://doi.org/10.1370/afm.2329

Conway, C. C., Forbes, M. K., Forbush, K. T., et al. (2019). A hierarchical taxonomy of psychopathology (HiTOP) for clinical practice. Clinical Psychological Science, 7(2), 236–259. https://doi.org/10.1177/2167702618810696

Du, Q., Ren, Y., Meng, Z., He, H., & Meng, S. (2025). The efficacy of rule-based versus large language model-based chatbots in alleviating symptoms of depression and anxiety: Systematic review and meta-analysis. Journal of Medical Internet Research, 27, e78186. https://doi.org/10.2196/78186

Forbes, M. K., Ringwald, W. R., Allen, T., et al. (2024). The principles and procedures underpinning the development of the Hierarchical Taxonomy of Psychopathology (HiTOP) model: Updating the model. Manuscript / HiTOP Revisions Workgroup.

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. Annual Review of Clinical Psychology, 12, 83–104. https://doi.org/10.1146/annurev-clinpsy-021815-093634

Gibbons, R. D., & deGruy, F. V. (2019). Without wasting a word: Extreme improvements in efficiency and accuracy using computerized adaptive testing for mental health disorders (CAT-MH). Current Psychiatry Reports, 21(8), 67. https://doi.org/10.1007/s11920-019-1053-9

Gibbons, R. D., Kupfer, D., Frank, E., Lahey, B. B., George-Milford, B. A., Biernesser, C. L., Porta, G., Moore, T. L., Kim, J. B., & Brent, D. A. (2020). Computerized adaptive tests for rapid and accurate assessment of psychopathology dimensions in youth. Journal of the American Academy of Child & Adolescent Psychiatry, 59(11), 1264–1273. https://doi.org/10.1016/j.jaac.2019.08.009

HiTOP Consortium. (2025). The Brief Hierarchical Taxonomy of Psychopathology (B-HiTOP): A 45-item self-report screening measure of six psychopathology spectra.

Hua, Y., Na, H., Li, Z., et al. (2025). A scoping review of large language models for generative tasks in mental health care. npj Digital Medicine, 8, 230. https://doi.org/10.1038/s41746-025-01611-4

Hua, Y., Siddals, S., Ma, Z., Galatzer-Levy, I., Xia, W., Hau, C., Na, H., Flathers, M., Linardon, J., Ayubcha, C., & Torous, J. (2025). Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: A systematic review. World Psychiatry, 24(3), 383–394. https://doi.org/10.1002/wps.21352

Kotov, R., Krueger, R. F., Watson, D., et al. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. Journal of Abnormal Psychology, 126(4), 454–477. https://doi.org/10.1037/abn0000258

Li, S. S., Balachandran, V., Feng, S., Ilgen, J. S., Pierson, E., Koh, P. W., & Tsvetkov, Y. (2024). MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In Advances in Neural Information Processing Systems 37 (NeurIPS 2024). arXiv:2406.00922

Lewis, C. C., Boyd, M., Puspitasari, A., et al. (2019). Implementing measurement-based care in behavioral health: A review. JAMA Psychiatry, 76(3), 324–335. https://doi.org/10.1001/jamapsychiatry.2018.3329

MedKGI. (2025). Iterative differential diagnosis with medical knowledge graphs and information-guided inquiring. arXiv:2512.24181

Q4Dx. (2026). A benchmark for evaluating diagnostic questioning efficiency of LLMs in patient conversations. Scientific Reports, 16. https://doi.org/10.1038/s41598-026-37022-y

Ringwald, W. R., Forbes, M. K., & Wright, A. G. C. (2023). Meta-analysis of structural evidence for the Hierarchical Taxonomy of Psychopathology (HiTOP) model. Psychological Medicine, 53(2), 533–546. https://doi.org/10.1017/S0033291721001902

Rony, M. K., Das, D. C., Khatun, M. T., et al. (2025). Artificial intelligence in psychiatry: A systematic review and meta-analysis of diagnostic and therapeutic efficacy. Digital Health, 11, 20552076251330528. https://doi.org/10.1177/20552076251330528

Ruggero, C. J., Kotov, R., Hopwood, C. J., et al. (2019). Integrating the Hierarchical Taxonomy of Psychopathology (HiTOP) into clinical practice. Journal of Consulting and Clinical Psychology, 87(12), 1069–1084. https://doi.org/10.1037/ccp0000452

Solhol, A., et al. (2008). Decision support in psychiatry: A comparison of a computerized structured clinical interview with manual diagnosis (CB-SCID1). [Decision support / structured clinical interview; sequential-reasoning caution].

U.S. Preventive Services Task Force. (2016). Screening for depression in adults: Recommendation statement. JAMA, 315(4), 380–387.

Wright, A. G. C., & Kotov, R. (2024). State of the science: The Hierarchical Taxonomy of Psychopathology (HiTOP). Behaviour Research and Therapy, 176, 104518. https://doi.org/10.1016/j.brat.2024.104518

Zhang, A., et al. (2021). A Bayesian approach for medical inquiry and disease inference in automated differential diagnosis. arXiv:2110.08393

Zhang, Q., Zhang, R., Xiong, Y., Sui, Y., Tong, C., & Lin, F.-H. (2025). Generative AI mental health chatbots as therapeutic tools: Systematic review and meta-analysis of their role in reducing mental health issues. Journal of Medical Internet Research, 27, e78238. https://doi.org/10.2196/78238