FACULTY OF PSYCHOLOGY AND EDUCATIONAL SCIENCES

ENHANCING TRANSLATIONAL ABILITIES OF INTENSIVE LONGITUDINAL DATA APPLICATIONS: AN ADAPTIVE APPROACH TO IDIOGRAPHIC MODELLING BY LEVERAGING LARGE LANGUAGE MODELS

Word count: 19.994

Stijn Van Severen
Student Number: 0210 9759

Supervisor(s): Prof. Dr. Geert Crombez & Dr. Annick De Paepe

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the Master's degree in Theoretical & Experimental Psychology

Academic year: 2024 – 2025

GHENT UNIVERSITY

Table of Contents

1. INTRODUCTION 1

1.1 Problem Statement and Objectives 1

1.2 Literature Study 3

1.2.1 Psychometric Modelling of Mental Health 3

1.2.2 Ontologies for Mental Healthcare 7

1.2.3 Large Language Models in Mental Healthcare 13

1.2.4 Development and Evaluation of Digital Interventions 20

1.3 Research Questions 26

2. METHODS 27

2.1 Development of the Software's Analytical Component 27

2.1.1 Development of PHOENIX Ontology 27

2.1.2 Development of Hierarchical Updating Algorithm 37

2.1.3 Development of Modular Agentic Framework 40

2.2 Evaluation of Software's Analytical Component 41

2.2.1 Evaluation of PHOENIX Ontology 41

2.2.2 Evaluation of Hierarchical Updating Algorithm 45

2.2.3 Evaluation of Modular Agentic Framework 46

3. RESULTS 52

4. DISCUSSION 53

5. CONCLUSION 54

6. AI ACKNOWLEDGEMENT SECTION 55

7. REFERENCES 56

8. APPENDIX 67

1. INTRODUCTION

1.1 Problem Statement and Objectives

The high prevalence and substantial economic burden of mental health disorders continue to pose a significant challenge for societies. In 2023, according to the Public Health Monitoring report (Zorgnet-Icuro, 2023), no fewer than 22% of the Belgian population experienced one or more mental disorders. The Organization for Economic Cooperation and Development estimated, in 2018, that mental health disorders incurred annual costs of €20.7 billion, equivalent to approximately 5% of Belgium's gross domestic product (OECD, 2018). Mental health, according to the World Health Organization (2022), can be conceptualized as follows:

> *"**Mental health** is a state of mental wellbeing that enables people to cope with the stresses of life, to realize their abilities, to learn well and work well, and to contribute to their communities. It is an integral component of health and well-being that underpins our individual and collective abilities to make decisions, build relationships and shape the world we live in. It is a basic human right, crucial to personal, community and socio-economic development. Mental health is more than the absence of mental disorders; it exists on a complex continuum, which is experienced differently from one person to the next, with varying degrees of difficulty and distress and potentially very different social and clinical outcomes."*

Under Belgium's federal e-Health action plan to strengthen electronic healthcare delivery, various programs for the development of evidence-based software solutions were subsidized (RIZIV, 2025). Despite many of these applications demonstrated their clinical effectiveness and efficacy, they still uniformly lack integrated (semi-)automated pipelines for real-time large-scale data acquisition, advanced analysis, and subsequent communication of these analysis results. This shortcoming is predominantly seen in digital interventions for mental health problems that require: 1) any form of longitudinal collection of multivariate time series for which sophisticated analyses and reasoning frameworks are needed to transform the data-driven insights into useful recommendations; 2) adaptive modeling procedures due to time-varying internal dynamics for which static group-level approaches are insufficient; and 3) an interoperable framework—with sufficient breadth and depth—to represent statistical associations between symptoms and potential solutions.

Therefore, the objective of this thesis is to develop a scalable software solution for digital interventions—by leveraging large language models—that facilitates the process of identifying treatment targets with the highest negative impact on the current mental state of an individual, followed by a theoretical translation into a clear actionable scheme for cognitive-behavioral change. Hulsmans et al. (2024) have shown that psychometric structures can differ markedly between participants; this between-person variability is frequently overlooked in clinical practice (Piccirillo et al., 2019; Kuper et al., 2024). As a result, a strong rationale has developed—from a treatment-oriented point of view—to move beyond conventional group-based (i.e., nomothetic) approaches when designing any type of digital intervention (Altman et al., 2020; Levinson et al., 2024).

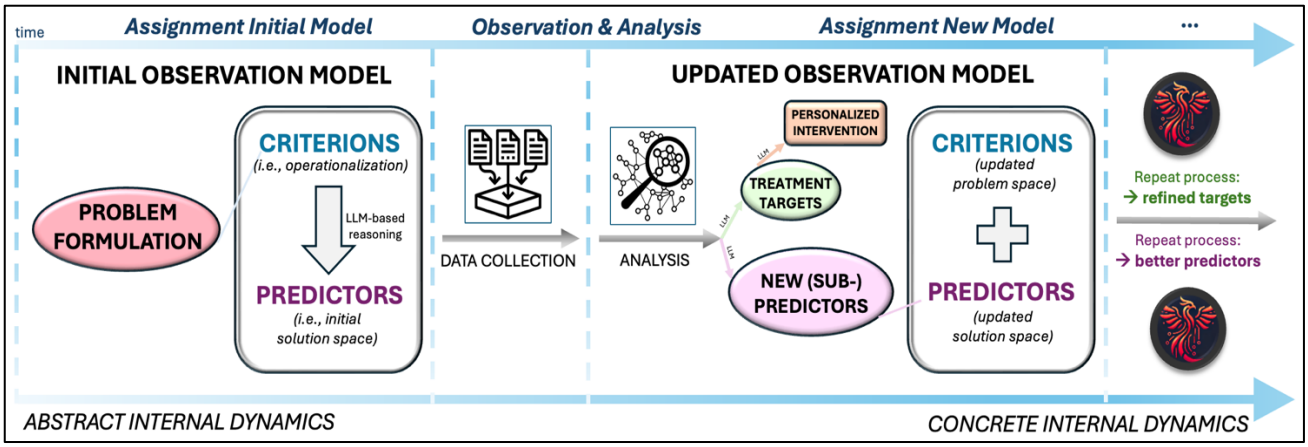
However, a fundamental issue with the current individual-level (i.e., idiographic) approach is that neither research nor clinical practice has yet succeeded in adequately translating network-analysis findings into a concrete personalized and contextualized action scheme (Hulsmans et al., 2024, Levinson et al., 2025). The absence of clear translational pathways primarily arises when, by design, important treatment variables—with the potential for ameliorating the psychopathological state of the client—are omitted in the observation model. Therefore, idiographic models should, at least, incorporate a set of biopsychosocial predictors that have the potential to statistically explain—and ultimately treat—a(ny) given mental state of an observed individual.

Motivated by this challenge, this master's thesis will be partially directing its focus to the construction of such a comprehensive formal representation of mental-well-being problems (i.e., criterion variables) and their corresponding action schemas (i.e., predictor variables). These will all be part of the **PHOENIX ontology**, which stands for '*Personalized Hierarchical Optimization Engine for Navigating Insightful eXplorations*'. This refers to the data-driven engine behind the software solution that operates by successively searching for treatment targets, defining mental health problems as statistically optimizable states within a large state space.

A key component of this optimization framework is the hierarchical updating (HU) algorithm, which operates on the PHOENIX ontology—allowing an initial representational model to progressively take on a more concrete form over time (Figure 1). This temporal component of adaptivity allows for a transition from a broad, abstract representation of internal dynamics to an in-depth, concrete representation. To furthermore facilitate the scalability of this application for digital intervention applications, large language models will be integrated directly into the architecture of the optimization engine. Their ability to handle extensive corpora of textual data—often reaching expert-level capacity—renders them as promising candidates for developing end-to-end services in mental healthcare (Lawrence et al., 2024; Guo et al., 2024; Hua et al., 2025).

Figure 1

Illustration of the Personalized Hierarchical Optimization Engine for Navigating Insightful eXplorations



Note. This optimization engine aims to tailor digital interventions by defining mental states as statistically optimizable problems for which a treatment-guided convergence of salutogenic satisfaction is possible.

1.2 Literature Study

1.2.1 Psychometric Modelling of Mental Health

1.2.1.1 Towards a Single-Subject Approach

A consensus among researchers and clinicians has been growing that traditional **nomothetic approaches**—which seek general laws by aggregating data across many people—often obscure crucial marked individual differences: *"what proves true at the group level may not apply to many individuals"* (Fisher et al., 2018; Beck & Jackson, 2019; Hulsmans et al., 2024). Awareness of this pressing issue increased significantly when researchers began systematically quantifying the extent of heterogeneity of mental disorders within diagnostic systems (Allsop et al., 2019; Zhao et al., 2025). For example, Olbert et al. (2014) found that in the majority of DSM-IV-TR and DSM-V, two individuals could receive the same diagnosis without sharing any common symptom—64% and 58.3%, respectively.

Therefore, this substantial between-subject variability within diagnostic classes—and the resulting problem of applying one-size-fits-all, group-derived treatments to clients with divergent symptom patterns—has driven calls for broader adoption of single-subject, **idiographic approaches** for modelling mental health-related processes (Molenaar, 2004; Wright & Woods, 2020; Schwarzbach et al., 2025). This approach offers a new potential to bridge the research-practice gap by focusing on the unique, within-person dynamics of a single individual over time (Hulsmans et al., 2024; Levinson et al., 2025). However, the widespread adoption of single-subject modelling has historically been constrained by the imposed logistical demands for the need of extended periods of intensive, longitudinal data collection (Frumkin et al., 2020; Zuidersma et al., 2024).

1.2.1.2 Analyzing Time Series using Vector Autoregression Model

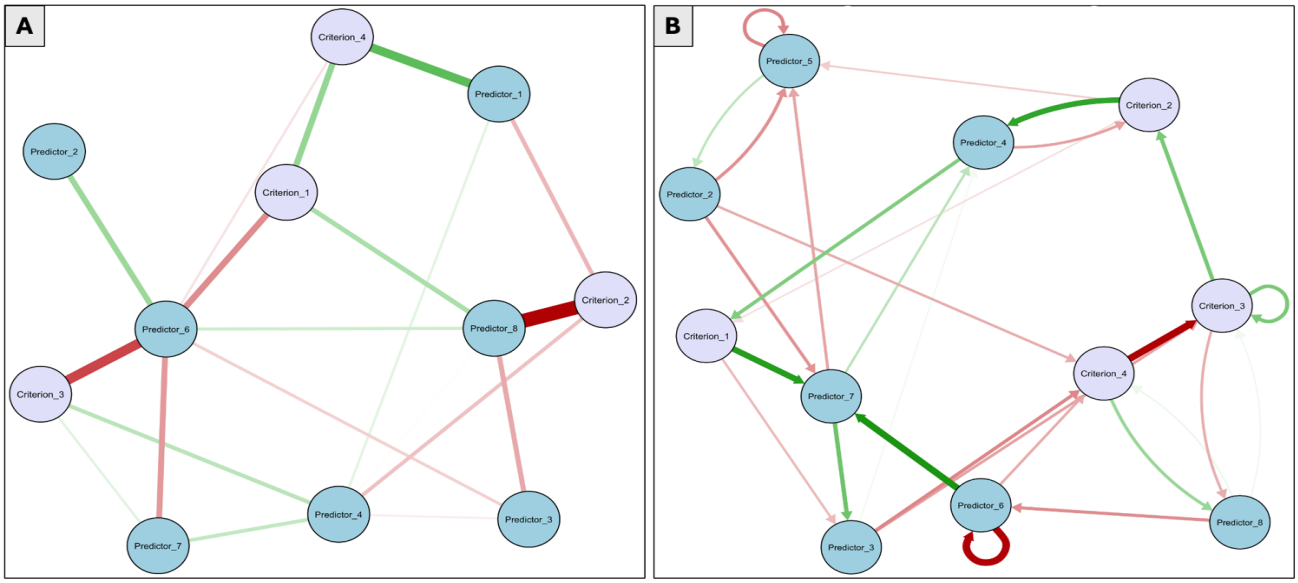
Fortunately, the digital proliferation of easy-to-use applications for ecological momentary assessments (EMA) and wearables-based data collection now allow intensive, real-time tracking of symptoms, behaviors, and physiological states at the individual level, laying the foundation for truly personalized case formulations and digital interventions (Henry et al., 2023; Levinson et al., 2024; Mink et al., 2025). To analyze the obtained longitudinal, multivariate time series data, graphical **vector autoregression models** (gVAR) are widely used in both fundamental and applied clinical research (Epskamp et al., 2018; Park et al., 2022; Levinson et al., 2025). This method operates by fitting a vector autoregressive model of a lag order of choice 'l' to the data.

During estimation, gVAR fits two interdependent components in one unified procedure: 1) a first component estimates a precision matrix of the residuals to reveal undirected, instantaneous associations among variables within the same assessment window; 2) a second component estimates a lag-one coefficient matrix whose directed entries quantify how each variable at time $t-l$ uniquely predicts every variable at time t. Both sets of parameters are regularized using the least absolute shrinkage and selection operator (LASSO) penalty—also known as the ℓ_1 penalty—whose strength is selected by minimizing an information criterion, thereby driving

negligible coefficients to zero and eliminating spurious connections (Vidaurre et al., 2013; McCulloch et al., 2024). By jointly optimizing these components, gVAR cleanly separates synchronous (i.e., *contemporaneous*) from predictive (i.e., *temporal*) relationships in a single, coherent modeling step, without requiring sequential or ad-hoc adjustments (Figure 2).

Figure 2

Contemporaneous and Temporal Network Estimates using Graphical Vector Autoregression



Note. Panel A depicts contemporaneous associations—variables that co-vary in the same measurement window after accounting for all lagged effects; Panel B depicts temporal links—how past values uniquely forecast future values. From an optimization point of view, these idiographic models can be used to narrow down any set of predictors (in blue) that could optimally change the values of the criteria (in purple).

The **contemporaneous network** emerges directly from the estimated residual precision matrix. After accounting for all lagged influences, the remaining covariance structure reflects instantaneous co-fluctuations among symptoms or physiological signals. These residual associations are inverted into a sparse precision matrix, where each nonzero off-diagonal element implies a unique partial correlation between two variables—i.e., a direct same-occasion dependency that cannot be explained by any other measured variable or by their past values. From a clinical perspective, this network reveals which symptom pairs tend to rise and fall together in real time, highlighting potential synchronous pathways (e.g., simultaneous spikes in anxiety and heart rate) that may warrant joint intervention. The regularization also mitigates the risk of identifying indirect or noise-driven associations, yielding a parsimonious graph whose topology can be probed for central 'hub' symptoms or clusters that co-activate within each assessment window.

By contrast, the **temporal network** is derived from the estimated lag-*l* coefficient matrix in the gVAR(*l*) component. Each row of this matrix corresponds to a regression of one outcome variable at time *t* on all

predictors at time $t-l$, producing directed edges—often called Granger-causal links—between prior and subsequent states. An edge from variable i to j indicates that past values of i carry unique predictive information about future values of j, over and above all other lagged variables. The LASSO regularization here suppresses weak or unstable cross-lagged effects, isolating only those predictive pathways that consistently improve out-of-sample forecast performance.

1.2.1.3 Handling Time-varying Networks

A major limitation of traditional vector autoregressive modeling in psychopathology is the assumption of stationarity—that the strength and structure of symptom-to-symptom associations remain constant over the observation period. In practice, many events—such as life stressors, or seasonal fluctuations in light density—can lead to non-stationary dynamics, causing static gVAR models to misrepresent both the timing and magnitude of key symptom interactions (Ryan et al., 2025). As Siepe et al. (2024) demonstrated, formally testing for stationarity with change-point procedures revealed pervasive non-stationarity in large multi-month multivariate time series; underscoring that average, time-invariant network estimates may obscure critical within-person shifts in symptom connectivity, or time-specific convergence to stable network connectivity.

To address this, a **time-varying gVAR** framework can be employed that leverages kernel-smoothed, locally weighted regressions to estimate a series of gVAR coefficient and precision matrices at several equidistant estimation points (Bringmann et al., 2024; Hoekstra et al., 2024). At each estimation point, observations are weighted according to their temporal proximity, so that data nearer in time exert greater influence on the local model fit. The width of this weighting window—known as the bandwidth—is chosen via a cross-validation procedure that balances the fidelity to local fluctuations against the stability of parameter estimates. Within each local gVAR fit, LASSO regularization is applied simultaneously to the precision matrix of residuals and the lag-l coefficient matrix. By jointly optimizing both matrices, this joint estimation procedure disentangles contemporaneous and temporal networks in a single coherent model—rather than treating them as separate, sequential steps—and thereby captures both gradual trends and abrupt shifts in symptom dynamics.

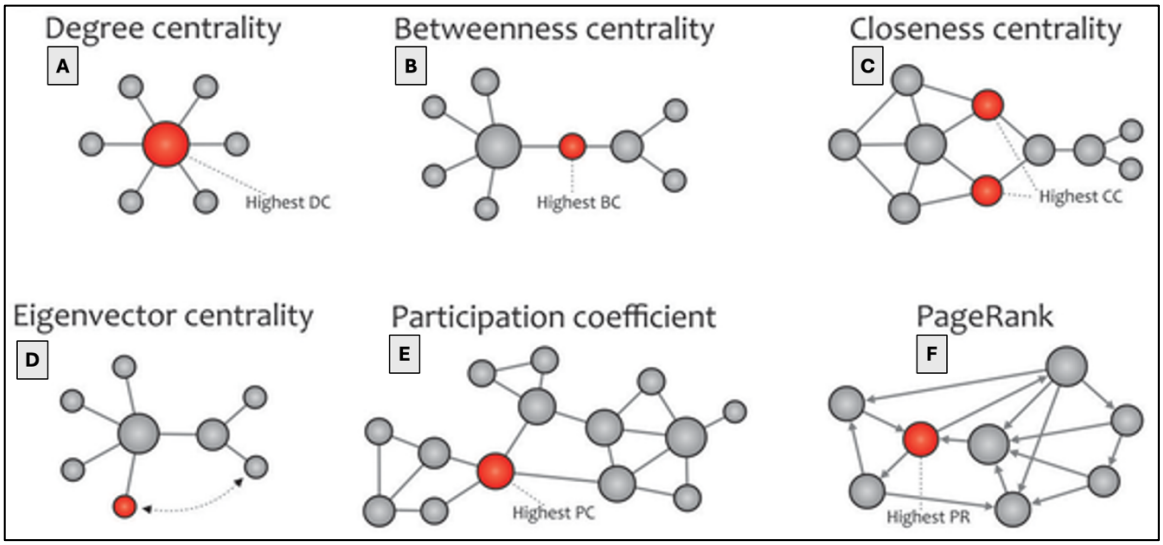
1.2.1.4 Analyzing Idiographic Networks using Graph Theory

Once the networks are obtained, candidate intervention points can be identified by evaluating symptom importance using **centrality measures** based on Graph Theory (Farahani et al., 2019; Serafim et al., 2025; Figure 3). To determine which centrality index most effectively alters network topology, Castro et al. (2024) applied two simulation-based attack procedures—normal attack, in which nodes were removed in descending order of their initial centrality scores, and cascade attack, in which centrality was recomputed after each removal to always target the most central remaining node—and at each iteration measured (1) the number of disconnected components, (2) the average shortest-path length, and (3) the network density once half the nodes had been eliminated. Their findings indicated that *degree centrality* yields the largest increases in

component count and path length alongside the steepest reductions in density. By ranking candidate variables based on centrality, clinicians can systematically prioritize symptoms and/or **treatment targets** whose modification is expected to produce the greatest network change—thereby guiding the personalized selection of high-impact intervention points (Kaiser & Laireiter, 2018; Lunansky et al. 2021; Serafim et al., 2025).

Figure 3

Local Network Centrality Metrics based on Graph Theory



Note. The following figure illustrates six commonly used local centrality metrics for identifying influential nodes within a network topology: A) Degree – the number of direct connections a node has; B) Betweenness – the proportion of shortest paths that pass through a node; C) Closeness – the inverse of the average shortest-path distance from the node to all others; D) Eigenvector centrality – a measure of influence that accounts for both a node's connections and the centrality of its neighbors; E) Participation coefficient – the extent to which a node's connections are distributed across different network modules; and F) PageRank – a variant of eigenvector centrality that balances the number and quality of incoming links, assigning higher scores to nodes connected to other influential nodes. This image is adapted from Farahani et al. (2019).

1.2.1.6 Ensuring Robust Network Estimation

An accurate estimation of psychopathological network parameters requires a rigorous evaluation of both the precision of edge weights and the stability of centrality indices under sampling variability. Nonparametric bootstrapping—typically involving 1,000 resamples with replacement—produces empirical distributions for each edge and centrality metric, allowing for the computation of 95% confidence intervals (Kim, 2014). Edges whose intervals exclude zero are deemed reliably nonzero, while nodes with narrow, non-overlapping strength or expected-influence intervals are interpreted as stable hubs rather than artifacts of random fluctuation (Epskamp et al., 2018; Borsboom et al., 2021). Case-dropping bootstraps further assess the robustness of centrality estimates by iteratively omitting increasing portions of the sample and recalculating centrality

metrics. The resulting correlation stability coefficient indicates the maximum proportion of cases that can be excluded while still preserving, for example, a correlation of ≥ 0.70 between the original and subset-based centrality rankings. Coefficients of ≥ 0.25 reflect acceptable stability, whereas values ≥ 0.50 point to robust hub identification despite considerable sample attrition (Epskamp et al., 2018; Christensen et al., 2024).

In addition to post-hoc robustness evaluation, simulation-based power analysis allows researchers to estimate the sample size necessary to detect key network features with adequate sensitivity before data collection (Epskamp & Fried, 2017; Borsboom et al., 2021). This approach involves specifying a plausible generating model—typically a Gaussian graphical model with predetermined sparsity, edge strengths (e.g., an edge of β = 0.20), and centrality patterns—and generating numerous synthetic datasets across a range of candidate sample sizes. Each dataset is analyzed using the same estimation methods and inferential thresholds intended for the target study. By calculating the proportion of simulations that successfully recover key parameters (e.g., an edge of β = 0.20), power curves can be constructed to visualize the relationship between sample size and detection probability with sufficiently high statistical power. This study-specific simulation-based strategy effectively generalizes classical power analysis to the high-dimensional, penalized estimation context of network psychometrics, while eschewing uniform sample-size targets (Christensen et al., 2024).

1.2.2 Ontologies for Mental Healthcare

1.2.2.1 What is an Ontology?

Without a properly formalized representation of the (non-)clinical state (i.e., *criterion variables*) and plausible therapeutic targets (i.e., *predictors*), the applicability of idiographic models remains limited: For if a mental health scenario is not adequately operationalized into a comprehensive set of criteria and plausible predictors, then selecting high-impact intervention targets would, by default, be epistemically sub-optimal. Relatedly, black box models—such as neural networks—are often criticized for lacking transparency of their internally hidden representations of complex state spaces (Buhrmester et al., 2019; Dobson, 2023). To address these limitations—of having inadequate and non-transparent representations—an interest in ontologies has been growing among researchers and clinicians to create systematic, and interoperable representations of mental-health related variables (Amoretti et al., 2019; Schenk et al., 2024; Zhu et al., 2024):

> *"An **ontology** is a formal, explicit specification of a shared conceptualization of a domain, expressed in a machine-interpretable language. Its structure typically comprises: (1) classes representing domain entities; (2) properties defining permissible relations between classes; (3) individuals instantiating the classes; and (4) axioms imposing logical constraints and enabling inference over classes, properties, and individuals."*

Ontologies achieve formal clarity and computational tractability by distinguishing between terminological knowledge and assertional knowledge: Terminological knowledge encodes the intentional structure of a domain by specifying its underlying conceptual schema. This is done by defining 1) *classes* (i.e., concepts), 2)

taxonomic hierarchies (i.e., recursive subclass relations), 3) *logical constraints*—such as disjointness axioms—, and 4) *semantic specifications for object properties*, including their domains, ranges, cardinality restrictions, and characteristics like transitivity, symmetry, reflexivity, and functional uniqueness. This layer provides the logical scaffolding for modeling domain semantics in a way that supports automated reasoning tasks such as consistency checking, and subsumption inference. Assertional knowledge, by contrast, constitutes the extensional layer of the ontology, grounding the abstract conceptual schema in concrete data. It consists of individual assertions that instantiate the classes and properties. These include: (1) *class assertions*, which specify that a given individual belongs to a particular class; (2) *property assertions*, which define relationships between individuals via object or data properties; and (3) *identity statements*, which establish sameness or distinction among individuals.

This architectural decoupling enhances modularity and maintainability, allowing domain models to evolve independently of individual datasets and supporting scalable, incremental reasoning workflows as schemas or data change. The explicit formalization of concepts and relationships also underpins semantic mappings between disparate ontologies, promoting reuse of core definitions across applications. Altogether, these features ensure that diagnostic categories, outcome measures, symptom descriptions, and personalized therapeutic interventions are defined unambiguously, remain interoperable across heterogeneous datasets, and are fully amenable to semantic querying, seamless data integration, and advanced knowledge discovery (Yamada et al., 2020; Belani et al., 2025).

1.2.2.2 Standardized Framework for Developing Ontologies

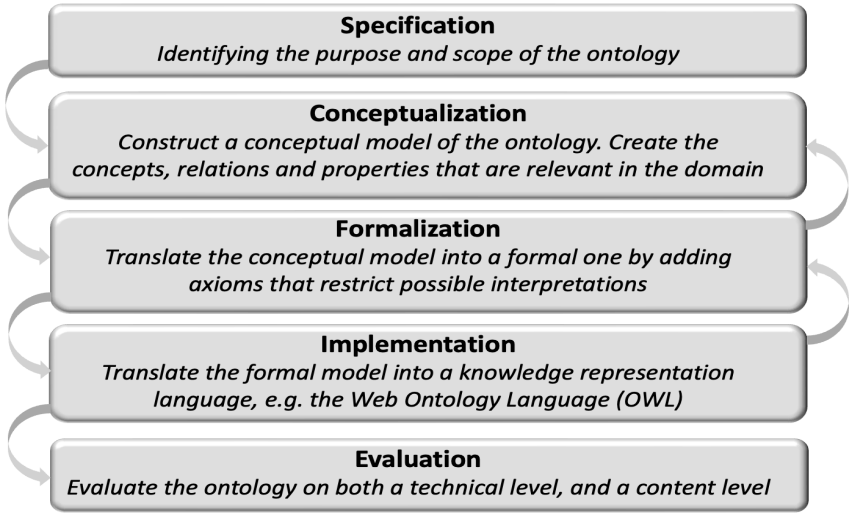
A prominent framework for developing ontologies has been constructed by Pinto and Martins (2004). Their methodology is structured into five interrelated components. First, 1) *specification* begins with a clear delineation of the ontology's intended purpose and scope—defining which mental-health constructs must be captured, the contexts in which the ontology will be applied, and communication with key stakeholders that will be involved in its development and use. Next, during 2) *conceptualization*, these identified constructs are translated into a high-level conceptual model. In this step, key classes and their relations are identified, and organized into a coherent schema that reflects the structural logic of the domain. The 3) *formalization* step then augments this model by introducing formal logical axioms—such as domain and range constraints, cardinality restrictions, and disjointness declarations—each of which contributes to the internal coherence of the ontology. These formal specifications serve to constrain possible interpretations of the ontology and enable automated deductive inference by ensuring semantic consistency across representational layers.

In 4) *implementation*, the formalized conceptual schema is encoded using a machine-readable ontology language. During this phase, the ontology's classes, properties, individuals, and axioms are systematically serialized for integration with semantic tools, and reasoning engines. Finally, 5) *evaluation* involves a dual validation process. Technical validation includes logical consistency checking, inferential completeness

assessment, and verification of adherence to community standards such as the OBO Foundry principles (Jackson et al., 2021; Braun et al., 2024; see Table 1 on page 41). Concurrently, content validation is conducted via expert review, ensuring that the ontology faithfully and exhaustively represents the intended mental-health domain. This often involves constructing and answering a representative set of competency questions (Figure 4) to empirically test the ontology's coverage and inferential adequacy.

Figure 4

Iterative Ontology Development Framework by Pinto and Martins (2004)



Note. Pinto and Martins (2004) propose a five-step ontology engineering framework.

1.2.2.3 An Overview of Modern State-of-the-art Ontologies of Mental Health

The majority of prominent ontology repositories tend to have a pronounced emphasis on biomedical semantics, which constrains their applicability for digital interventions that aim to incorporate cognitive-behavioral dimensions into their decision-support systems (BioPortal, 2024; GALENOS, 2024; Ontobee, 2024). Nevertheless, Larsen and Hastings (2018) have initiated a formal ontological approach specifically tailored to mental health, aiming to represent its complex affective, cognitive, and diagnostic dimensions in a structured, interoperable framework. Their tripartite ontology serves as an integrative semantic framework specifically designed to bridge the disciplinary gap between affective science and psychiatric diagnostics. It accomplishes this by formally connecting three modular ontologies: 1) the *Emotion Ontology*, which has the ability to capture affective entities such as emotions, moods, and associated physiological and behavioral markers; the 2) *Mental Functioning Ontology*, which models non-pathological cognitive and experiential processes such as perception, memory, and belief; and 3) the *Mental Disease Ontology*, which systematically codifies mental health disorders and their associated symptomatology.

The architecture of this mental health ontology supports bridging axioms—logical relationships that explicitly relate affective phenomena to psychiatric symptom and diagnostic—while remaining theoretically agnostic to

specific explanatory models. By enabling semantic alignment between 1) dimensional constructs—such as those used in RDoC framework (Cuthbert et al., 2020; Morris et al., 2022), 2) clinical categories—such as those used in the DSM-V-TR (American Psychological Association, 2023) or ICD-11 (World Health Organization, 2024)—, 3) and computational annotations (e.g., used in biomedical ontologies), this tripartite ontology framework facilitates cross-disciplinary data integration, automated reasoning, and consistent annotation of empirical findings across diverse research paradigms. Consequently, their approach seems to be particularly suited for accommodating the biopsychosocial complexity of mental disorders, offering a flexible but formally rigorous infrastructure for synthesizing evidence, supporting transdiagnostic modeling, and refining psychiatric classification schemes in light of emerging data.

Schenk et al. (2024) further bridged this semantic gap of mental health ontologies, conducted under the auspices of the GALENOS project. Their primary objective was to tackle the pervasive issue of inconsistent terminology and fragmented research across the following three highly prevalent mental disorders—anxiety, depression, and psychosis. By constructing an ontology designed to structure and integrate evidence from a series of living systematic reviews, their project aimed to enhance evidence synthesis and accelerate the advancement of more effective preventive and therapeutic interventions. The resulting ontology serves both as a classification framework and as a tool for the semantic organization of research data in an accessible, structured online repository. However, a notable limitation of both aforementioned mental health-related ontologies concerns their constrained semantic scope, which can be partially attributed to methodological design choices that ultimately hinder the scalability of their formalized representation of mental health.

1.2.2.4 Dealing with Scalability Issues during Ontology Development

Zhu et al. (2024) tackled this critical bottleneck of scaling knowledge graph construction by harnessing the generative power of large language models (LLM). They operationalized an explicit ontology schema as structured ‘competency questions’, which guide prompt templates that target specific classes and relations. These prompts were executed in parallel across large batches of unstructured text, with an automated prompt-refinement module that analyzed initial LLM outputs, adjusted template parameters, and re-issued queries to maximize alignment with the ontology’s controlled vocabulary. By exploiting commodity GPU clusters for concurrent LLM invocations and employing a schema-aware filtering layer to discard out-of-vocabulary or low-confidence triples, their pipeline sustained processing rates in the order of 10^5 abstracts per day, while preserving over 90 % precision against a held-out, manually annotated gold standard.

Building on this extraction backbone, they layered a closed-loop validation mechanism: the partially instantiated graph is used to generate targeted verification prompts that cross-check edge consistency and detect missing links. Schema-derived constraints triggered automated re-prompting for any discrepancies, and a lightweight human-in-the-loop review flags ambiguous subgraphs for expert adjudication. To generalize beyond fixed pipelines, Zhu et al. further introduced **AutoKG**, a modular multi-agent framework in which

distinct LLM agents assume roles—Extractor, Validator, and Retriever—and communicate via a task queue. 1) The *Retriever* agent interfaces with external knowledge sources to ground extracted assertions; 2) the *Validator* iteratively refines the graph against ontology axioms; and the 3) *Extractor* focuses on novel entity and relation instantiation. This orchestrated, schema-driven multi-agent design not only automates continuous updates to the knowledge graph but also ensures high semantic fidelity, providing a scalable foundation for dynamically evolving domain ontologies.

1.2.2.5 Machine Learning Approach to Remove Overlapping Structures

To further enhance—and often reverse the induced 'damage' by an LLM-based graph construction—the overall applicability of ontologies during real-time inference, *semantic redundancy* should be minimized to optimize latency and reduce computational costs. To investigate this, Forbes et al. (2023) constructed a knowledge graph of mental health where the full class structure was obtained by extracting domain, disorder and symptom names from the Diagnostic and Statistical Manual of Mental Disorders (American Psychological Associated, 2023). The authors first harvested all 202 adult-focused diagnoses and specifiers from DSM-V section II, then parsed their textual criteria into 1419 constituent symptom statements by splitting disjunctive clauses (e.g., 'insomnia' or 'hypersomnia') into separate entries.

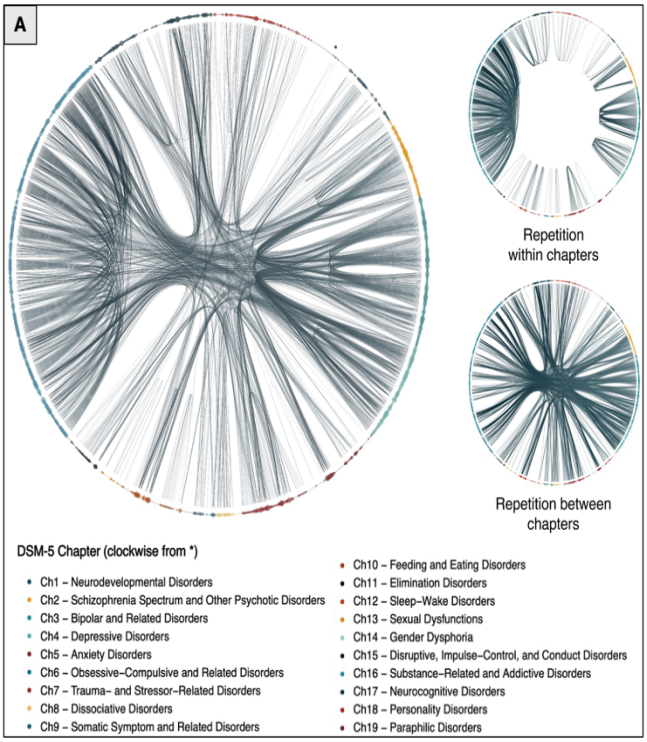
Next, they performed a two-stage redundancy analysis: (1) manual content coding, where four trained researchers independently assigned each statement to one of four super-categories (1. affective; 2. cognitive; 3. behavioral; and 4. somatic), subdivided these into finer-grained semantic clusters, and iteratively merged any statements judged to capture the same subjective experience; and (2) neural semantic matching, in which they fine-tuned a transformer encoder on 1067 hand-labeled pairs ('definitely same' vs. 'definitely different') and then scored the remaining ~566.000 possible symptom pairs ($AUC = 0.859$; under five-fold cross-validation). The top 1000 highest-scoring pairs were then manually adjudicated for any additional matches, yielding a final set of 3096 semantically redundant pairings.

From this pipeline they distilled the original 1419 statements into 628 distinct symptoms, a 56% reduction in redundant terms (Figure 5). Of these, 397 (63.2%) were unique—appearing in only one diagnosis—while 231 (36.8%) recurred across multiple diagnoses, totaling 1022 repeats (median = 3, range 2–22). Chapter-level analysis revealed that 140 of the 202 diagnoses (69.3%) included at least one non-unique symptom; entire chapters such as bipolar and related disorders, trauma- and stressor-related disorders, dissociative disorders, neurocognitive disorders and personality disorders exhibited 100% within-chapter repetition, whereas elimination disorders, gender dysphoria and paraphilic disorders showed zero overlap. The most pervasive cross-diagnostic symptoms—insomnia (in 22 diagnoses), difficulty concentrating (17), hypersomnia (17) and irritability (16)—were overwhelmingly drawn from major depressive disorder criteria, underscoring how symptom redundancy can obscure disorder-specific signals and inflate apparent comorbidity.

Figure 5

Semantic Redundancy in Diagnostic and Statistical Manual of Mental Disorders

A



B

Symptom	Number of diagnoses	Number of chapters
*Insomnia	22	8
*Difficulty concentrating	17	7
*Hypersomnia/sleepiness	17	6
Irritable mood	16	8
*Psychomotor agitation	16	5
*Depressed mood	15	6
*Psychomotor retardation	14	5
Hallucinations	14	5
*Fatigue	12	6
*Increased appetite	12	5
Anxiety	12	4
Restlessness	10	6
*Weight loss	10	5
Euphoria	10	4
*Decreased appetite	10	4

Note. A) Each dot on the circumference represents one of the 1419 constituent DSM-5 symptoms; dot size reflects symptom frequency across 202 primary disorders and specifiers. Lines connect symptoms that repeat within or between diagnostic chapters. B) Symptoms are sorted by the number of diagnoses in which they occur. The most pervasive cross-diagnostic symptom was 'insomnia'—in 22 diagnoses across eight chapters.

1.2.2.6 Implementing Ontologies in Semantic Recommendation Systems

Once a concept—such as 'mental health'—has been properly formalized into a non-redundant ontology, it can be employed in recommendation systems that use pre-defined behavioral change techniques with the ultimate goal in mind of alleviating mental health problems (Larsen et al., 2016). A recently developed ontology, by Braun et al. (2024), is the 'COPPER ontology' that integrates multiple sub-ontologies to increase the likelihood of behavioral change. These sub-ontologies, from a high-level overview, include: a) *Profile* sub-ontology where entities collectively represent potential user profiles; b) *Context* sub-ontology that formalizes context-relevant items to, in conjunction with the first ontology, enables personalization and contextualization of the recommendation system; c) *Barrier* sub-ontology where relevant barriers can be mapped onto specific recommendations to better prepare the user for potential behavioral or cognitive obstructions; and d) *Coping* sub-ontology that can aid in dealing these identified obstructions. The structure of this ontology theoretically aligns with the Health Action Process Approach (Schwarzer & Luszczynska, 2008) that explicitly conceives the adoption, initiation and maintenance of health behaviors as a componential process—including action plans and coping plans with barriers and resources.

By leveraging ontology-driven semantic reasoning, recommendation engines can move beyond explicit keyword matching, or black-box machine learning-based approaches to deliver contextually relevant and precise suggestions (Al Khatlib et al., 2024; Belani et al., 2025). In practice, this allows the system to interpret a user's symptom profile, treatment history, and environmental context as interconnected concepts—each defined by formal classes, properties, and axioms—so that recommended interventions are not just statistically correlated, but semantically appropriate. In mental healthcare, ontologies encode clinical guidelines and evidence-based behavior-change strategies, enabling a reasoner to infer care pathways (Luschi et al., 2023).

1.2.3 Large Language Models in Mental Healthcare

1.2.3.1 What is a Large Language Model?

In the rapidly evolving landscape of artificial intelligence, state-of-the-art large language models from organizations such as OpenAI, Meta and Google have been able to revolutionize various industries within a short timeframe (Raiaan et al., 2023). Their encoded knowledge, capacity to process extensive contextual information, and generative personalization capabilities render LLM's as promising candidates for developing end-to-end (i.e., fully automated) services in mental healthcare (Guo et al., 2024; Hua et al., 2025):

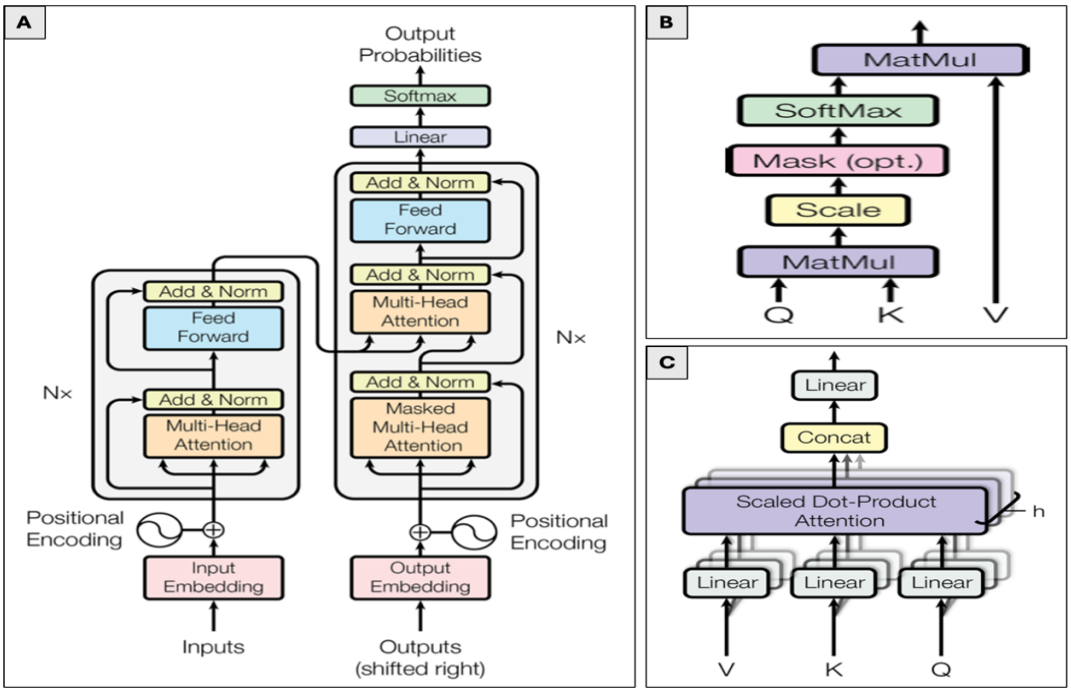
> *"A **large language model** is a deep neural network—commonly implemented using a transformer architecture—comprising stacked layers of multi-head self-attention mechanisms and feed-forward networks. It is trained on mass-scale datasets of text to model the conditional probability of each token given its preceding context, capturing high-dimensional statistical dependencies across sequences of tokens. Through autoregressive or masked token prediction objectives, the model learns complex syntactic, semantic, and contextual representations, enabling tasks such as text generation, infilling, summarization, translation, reasoning, and question answering. Often, they undergo a second phase of training using reinforcement learning from human feedback, which—typically ethically and pragmatically—aligns the model's output with human preferences by optimizing responses based on ranked comparisons or explicit reward signals."*

Nearly all modern large language models are built on the *Transformer architecture*—introduced by Vaswani et al. (2017), which is expected to become one of the most-cited papers of this century. This architecture is structurally organized as a series of identical stacked layers, each combining multi-head self-attention—implemented via multiple parallel scaled dot-product attention heads that linearly project input embeddings into query, key, and value vectors—with position-wise feed-forward networks composed of two dense layers separated by an intermediate dimensionality expansion and non-linear activation. Within each layer, every token can attend to every other token in the sequence through scaled dot-product attention, capturing both short- and long-range dependencies; additional dropout procedures applied to the attention weights and to the feed-forward outputs regularize these computations. Each attention and feed-forward sub-layer is wrapped in a residual connection to preserve gradient flow during deep training, and followed by layer normalization to stabilize volatile distributions of hidden states. To preserve sequence order information, learned positional

embeddings are combined with the input token embeddings, providing either absolute or relative positional cues—compensating for the attention mechanism's inherent permutation invariance (Figure 6).

Figure 6

Original Transformer Architecture – "Attention is All You Need" (Vaswani et al., 2017)



Note. A) A schematic of the original Transformer showing an encoder stack of *N* identical layers—each with a multi-head self-attention sub-layer and a position-wise feed-forward sub-layer and a decoder stack of *N* layers that add a masked multi-head self-attention sub-layer plus an encoder–decoder cross-attention sub-layer before the feed-forward block. Learned positional embeddings are added to input token embeddings at both encoder and decoder to inject sequence order information. B) Scaled dot-product attention computes the output as a weighted sum of value vectors, where weights are the softmax of the dot products between queries and keys divided by $\sqrt{d_k}$ C) Multi-head attention applies several parallel scaled dot-product attentions with independent projections, then concatenates and linearly transforms their outputs.

Topological variants of this original Transformer framework rewire or mask self-attention to serve different modeling objectives: a) In the *encoder-only* design, self-attention is bidirectional, and models are trained via masked-token prediction objectives to generate deep contextual embeddings optimized for understanding and classification tasks (Kamath et al., 2022). b) The *decoder-only* variant employs a causal mask so each token may only attend to preceding ones; with a next-token prediction loss, these models are specialized for free-form text generation (Radford et al., 2019). c) *Encoder–decoder* architectures combine a bidirectional encoder with a causal decoder linked by cross-attention, and are trained on sequence-to-sequence objectives—such as denoising, masked token prediction, or next-sentence generation—that enable conditional generation tasks like translation, summarization, and question answering (Ferraris et al., 2025).

1.2.3.2 Modern Enhancement Techniques for Using Large Language Models

Sparse *Mixture-of-Experts* (MoE) architectures further enhance model capacity by splitting the standard feed-forward layer into a large pool of independent expert subnetworks and inserting a lightweight gating network ahead of them (Liu et al., 2025). For each input token, the gating network computes a score for every expert—typically via a shallow projection of the token embedding followed by a sparsity-encouraging top-k selection—and dispatches the token only to the highest-scoring experts. Each selected expert processes the token through its own feed-forward computation in parallel, after which their outputs are aggregated (often by weighted sum) and passed onward. Because only a small subset of experts is activated for each token, the total parameter count can scale into the hundreds of billions while the floating-point operations per token remain roughly constant. In practice, effective MoE implementations incorporate auxiliary loss terms to encourage balanced expert utilization, optimize communication primitives to batch-route tokens across accelerators with minimal latency, and shard expert weights strategically across devices to prevent memory fragmentation and network congestion (Fedus et al., 2021; Du et al., 2024; Sukhbaatar et al. 2024 Sankar & Dimitri, 2025).

For specialized applications, *fine-tuning* a pretrained Transformer—by backpropagating gradients through every model parameter on task-specific data—remains the canonical adaptation strategy. However, this approach incurs substantial compute, and risks overwriting valuable general representations. Parameter-efficient fine-tuning addresses these challenges by freezing the backbone and introducing a minimal set of trainable components (Xu et al., 2023; Han et al., 2024). Low-rank adaptation decomposes each dense weight update into two low-rank matrices that are injected into the Q, K, V, and feed-forward projections, while adapter modules insert a narrow bottleneck—down-projection, nonlinearity, up-projection—between self-attention and multi-layer perceptron sub-layers (Hayou et al., 2024). Alternative methods such as BitFit adjust only bias terms, and prefix-tuning optimizes a small sequence of continuous prompt embeddings prepended to the input (Zhang et al., 2025). By cutting tunable parameters from $O(d^2)$—indicating quadratic complexity in model size—to $O(r \cdot d)$—a reduced linear complexity with rank factor r—these techniques slash fine-tuning computational cost and storage by three to four orders of magnitude, enable instant rollback via adapter removal, and facilitate rapid multi-task deployment through lightweight adapter swapping (Hu et al., 2021).

Finally, tool-enabled extensions upgrade a text-only Transformer into an **agentic AI platform** by integrating dedicated modules for perception, planning, and action execution (Liu et al., 2025; Acharya et al., 2025). In the *perception* pipeline, raw images can be partitioned into uniform patches, each projected into high-dimensional embeddings that are injected into the Transformer via specialized cross-attention layers equipped with learned modality-alignment matrices and gating scalars. For *planning* and *control*, the Transformer's final-layer representations are routed into lightweight 'planner heads that translate latent features into continuous spatial trajectories or discrete motor commands compatible with robotics middleware. External tools and services are accessed through a schema-aware function-calling interface: reserved tokens mark function identifiers and parameter slots, and a constrained output head serializes the model's calls into strict

drive particular outputs (Owen et al., 2024); and c) *ethical concerns* stemming from the potential regurgitation of memorized sensitive data and the inability to enforce functional constraints once the large language model is open-sourced (Alkamli et al., 2024; Liu et al., 2025).

1.2.3.4 Controlling Generative Outputs through Retrieval Augmented Generation

While studies have demonstrated that increasing the LLM's context window can reduce the likelihood of unfaithful hallucinations (Arif et al., 2023), such hallucinations remain a significant impediment to the reliable deployment of end-to-end LLM-based services—especially in high-stakes or domain-sensitive contexts where factual precision is non-negotiable. The limitation arises because even with an expanded context window, the model's reasoning remains tethered to its pretrained knowledge and token-limited working memory, both of which can lead to outdated, irrelevant, or fabricated information. To mitigate this challenging obstruction, **retrieval-augmented generation** (RAG) architectures can be employed, wherein an LLM's generative capabilities are dynamically grounded in external, domain-specific context retrieved from a vector-indexed database. This approach alleviates reliance on the model's internal parameters alone by incorporating up-to-date and task-relevant information at inference time (Jiang et al., 2023; Gao et al., 2023).

Building on this paradigm, Akkiraju et al. (2024) introduced a comprehensive RAG-based architecture explicitly designed to curb spurious content generation. Their *FACTS* framework—encompassing content Freshness, flexible Architectures, cost Economics, rigorous Testing, and robust Security—orchestrates semantic embedding fine-tuning, vector-database retrieval enhanced by query rephrasing and reranking, prompt engineering that enforces document access controls, and concise, reference-aware response synthesis. By defining fifteen control points across the RAG pipeline, this approach systematically mitigated hallucination risk while maintaining delivery of enterprise-grade performance.

Edge et al. (2024) furthermore observed that, although conventional RAG-based architectures (Lewis et al., 2020) reliably ground an LLM's generations in locally retrieved document passages, they lack the capacity to synthesize global thematic insights across an entire corpus, rendering them ineffective for query-focused summarization (Dang, 2006). To address this, they proposed a *Graph-RAG* pipeline in which an LLM first extracts entities and relations from source texts to construct a weighted entity knowledge graph, then applies hierarchical community detection using the Leiden algorithm (Traag et al., 2019) to partition this graph into coherently clustered subgraphs. Each community is then abstractively summarized via domain-tailored LLM prompts, yielding a set of 'community summaries' that serve as inputs to a map-reduce style generation—in which partial answers are produced in parallel for each summary and subsequently merged into a final, global response. In empirical evaluations on corpora of roughly one million tokens, Graph RAG consistently outperformed a naïve vector-search RAG baseline—achieving win-rate gains of up to 31% in diversity and up to 29% in comprehensiveness—, thereby demonstrating scalable performance for true global sense-making while preserving the local grounding essential for factual accuracy.

1.2.3.5 Creating Interoperable High-Dimensional Representations of Unstructured Data

A core component in all these types of textual RAG-based systems—and large language models in general—are embeddings (Tang et al., 2024; Petukhova et al., 2024; Rayhan & Ashrafuzzaman, 2025):

> *"An **embedding** is a learned, continuous vector representation of discrete data—such as words, tokens, sentences, or entire documents—constructed via a neural network to encode semantic, syntactic, or structural properties of the input. It maps symbolic inputs from a high-dimensional, sparse space into a dense, low- to mid-dimensional vector space, where geometric relationships reflect meaningful similarities. Embeddings are typically optimized through self-supervised, contrastive, or predictive learning objectives, depending on the architecture and training task. The resulting vector space is structured such that semantically similar inputs cluster together under a chosen similarity metric—commonly cosine similarity—, thereby enabling downstream tasks like classification, retrieval, clustering, or reasoning in a numerically tractable form."*

A powerful application of embeddings lies in constructing *vector databases* of large-scale patient records where millions of electronic health records can be collapsed into dense, fixed-length vectors that supported sub-second similarity search, clustering, and downstream analytics (Huang et al., 2021). Steiger and Kroll (2023) demonstrated this with Pat2Vec, a paragraph-based framework that transforms each patient's complete outpatient diagnosis profile into 10 to 100-dimensional vectors—by using Bayesian tuning of window size, negative-sampling exponent, learning rate, number of epochs, and choice of distributed memory versus distributed bag of words. They then evaluated performance on four clinical prediction tasks—case counts, emergency visits, age, and gender. Pat2Vec outperformed one-hot baselines, remained robust under 50% code dropout, generalized to much smaller cohorts, and uncovered clinically meaningful patient sub-cohorts.

These findings were corroborated by Rijcken et al. (2025), whose large-scale evaluation of clinical document classification revealed—for the first time in decades—that transformer-derived embeddings significantly outperform traditional rule-based systems. The authors attributed this divergence to the inability of symbolic methods to generalize across unstructured narratives, whereas transformer-based models captured latent semantic relationships through self-supervised pretraining. Together these findings suggest that these newer embedding-based methods offer a unique combination of efficient storage, easy access for subsequent analyses, and rapid similarity searches across millions of patient records enabling reliable, low-latency RAG-based solutions—such as conversational agents (Wang et al., 2024; Casella & Wang, 2025).

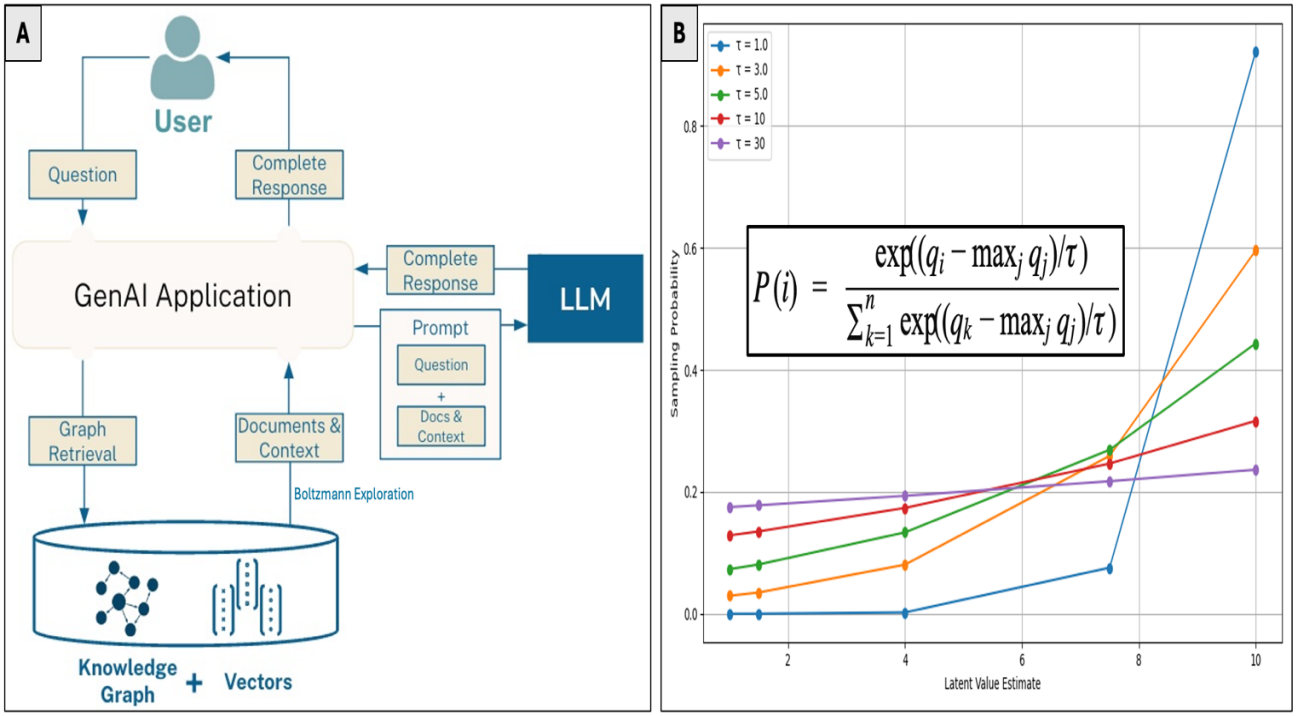
1.2.3.6 Balancing between Exploitation and Exploration during Medical Reasoning

In traditional RAG-based architectures, retrieval proceeds by greedily selecting the set with the highest embedding-based similarity scores to a given query. By contrast, the *Boltzmann exploration policy*—adapted from reinforcement learning frameworks (Cesa-Bianchi et al., 2017)—casts retrieval as stochastic sampling rather than pure maximization. Each candidate's similarity score is exponentiated after being scaling by a

temperature hyperparameter, and retrieval candidates are uniformly normalized to form a single probability distribution. Sampling from this distribution allows retrieval candidates with low to moderate relevance scores to be retrieved with non-zero probability (i.e., nevertheless given a chance to be selected), thereby promoting thematic diversity and reducing the risk of overlooking sub-optimal, yet useful, information (Figure 7).

Figure 7

Illustration of Graph-based Retrieval Augmented Generation with Boltzmann Exploration Policy



Note. A) Panel A depicts the overall architecture in which relevant context is retrieved from a knowledge graph to enrich the user prompt before generating a response; adapted from Rathle (2024). Panel B illustrates the impact of temperature *'τ'* on the sampling probability during the selection of relevant information. Higher temperature values equalize the (numerically stabilized) Boltzmann sampling probabilities. This can become advantageous for medical software solutions where overly greedy selection must be avoided to ensure that items with marginal latent relevance *'q'*, yet sufficient clinical significance, are not systematically overlooked.

Balancing exploitation and exploration via temperature-scaled softmax sampling can be critical for clinical RAG-based pipelines. By exponentiating and normalizing embedding affinities into a Boltzmann distribution, the system preserves top-ranked passages while allocating non-zero probability mass to mid-rank candidates that capture idiosyncratic symptom expressions, low-prevalence phenotypes, and atypical case modalities. This controlled stochasticity mitigates dominant-pattern overfitting and dataset biases, ensuring long-tail patient presentations remain accessible for downstream inference and facilitating highly granular, data-driven personalization of diagnostic and therapeutic recommendations (Renze et al., 2024; Xuan et al., 2025)

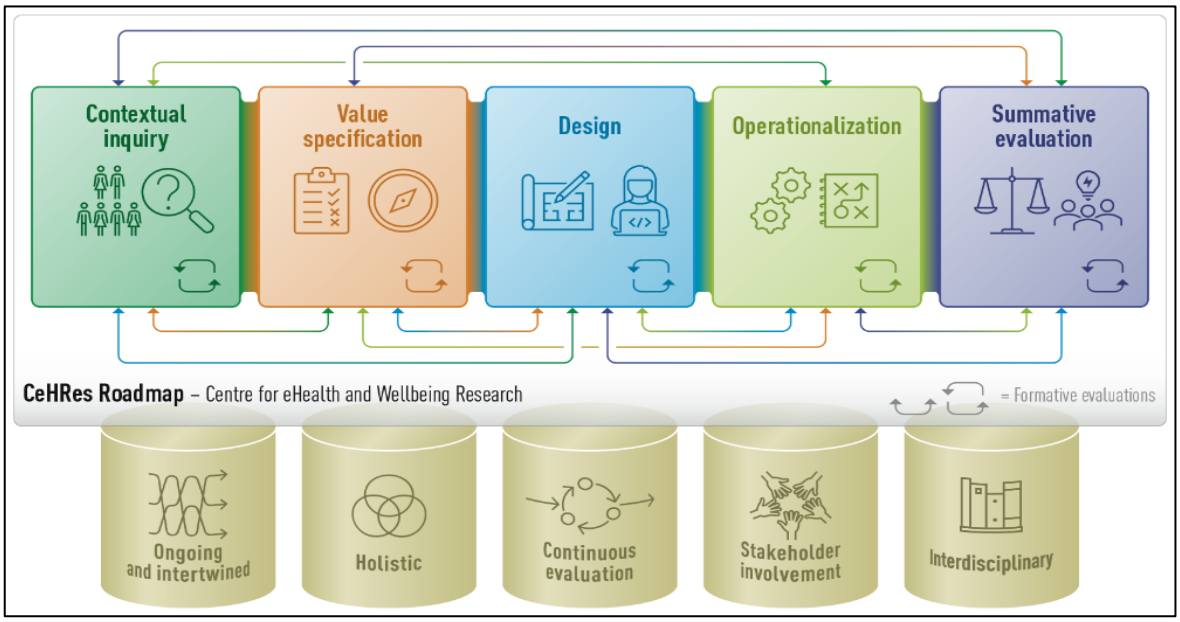
1.2.4 Development and Evaluation of Digital Interventions

1.2.4.1 Standardized Framework for Developing Digital Interventions

With the rapid advancement of promising (abovementioned) technologies for mental healthcare, the demand for a standardized framework—to steer their implementation within practice has never been more urgent. To ensure that an eHealth technology fits with its intended users, and the technical context within which it will be used, Kip et al. (2024) created the **CeHRes Roadmap**. Unlike other frameworks, this roadmap is explicitly non-sequential and emphasizes the dynamic interplay between its phases. It serves not as a rigid protocol, but as a guiding structure rooted in key design principles, enabling the tailoring of methods and activities to project-specific needs, stakeholders, and contexts. The model distinguishes five core phases: (Figure 8):

Figure 8

The CeHRes Roadmap 2.0 – Centre for eHealth and Wellbeing Research



Note. This figure is adapted from Kip et al. (2024) and illustrates the underlying five pillars of their roadmap.

1) *Contextual inquiry* seeks to construct a detailed model of the socio-technical ecosystem in which the intervention will operate. It involves systematic stakeholder mapping, ethnographic observation of end-user behaviors and workflows, and rigorous environmental scanning to uncover operational constraints and points of leverage. The outputs of this phase—comprehensive stakeholder profiles, contextual use cases, and gap analyses—form the empirical and theoretical substrate for all subsequent development activities.

2) Building on this, *value specification* converts these insights from the contextual inquiry into prioritized stakeholder objectives and quantifiable system requirements. This entails eliciting and ranking stakeholder imperatives (e.g., patient autonomy), selecting appropriate behavior-modification strategies grounded in psychological theory, and drafting a viable economic model to ensure long-term sustainability.

3) Design is an iterative cycle of prototype generation, stakeholder-driven refinement, and usability assessment. Beginning with low-fidelity mock-ups and advancing to high-fidelity interactive models, design teams integrate content modules, behavior-change components, and user-interface elements. Prototypes are subjected to structured usability testing—such as task-based walkthroughs and heuristic reviews—and then co-created with end users to verify that each iteration remains aligned with contextual requirements and can scale effectively in real-world settings.

4) *Operationalization* encompasses the deployment and embedding of the technology into its intended organizational and cultural milieu. Activities include finalizing the economic framework, diagnosing and mitigating barriers to adoption (e.g., regulatory constraints or workflow misalignments), and formulating a comprehensive implementation plan that articulates training protocols, support processes, and dissemination strategies. This phase ensures that the solution transitions from prototype to practice with minimal friction.

5) Summative evaluation assesses the intervention's performance against its predefined objectives by examining adoption metrics, clinical and organizational outcomes, and the mechanisms driving observed effects. By employing both quantitative and qualitative methodologies—such as controlled trials, usage-log analysis, and in-depth interviews—this evaluation phase illuminates not only whether the technology achieves its targets, but also elucidates the contextual and behavioral processes that underpin its success or failure.

6) Finally, the 'last phase' is *formative evaluation*, which interconnects all phases by providing continuous feedback loops to enhance both process and product. This happens through real-time data collection, stakeholder review sessions, and iterative verification of phase outputs; enables proactive identification of emergent issues and timely adaptation of both strategic and technical components to maintain alignment with the project's overarching aims and practical constraints.

1.2.4.2 The Importance and Methods of Personalizing Digital Interventions

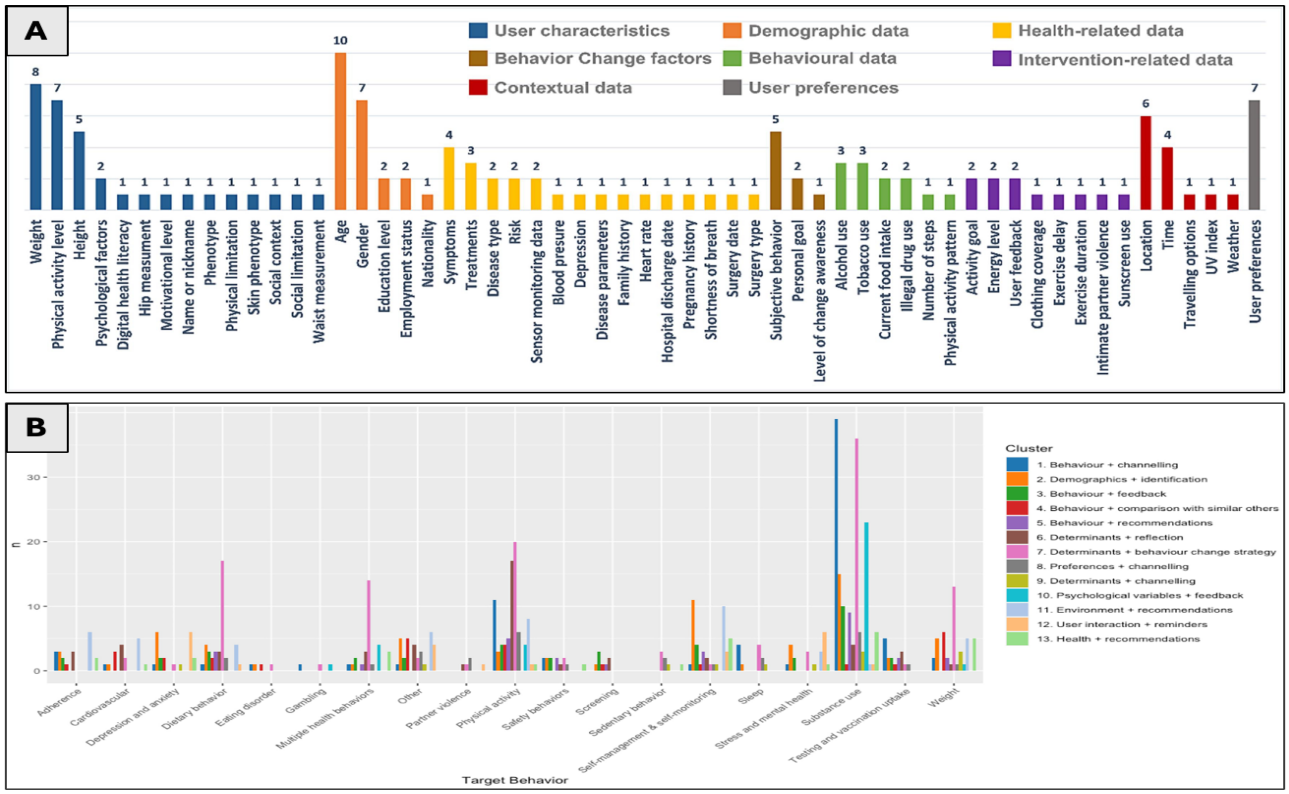
During all the abovementioned phases—but predominantly noticeable during the value specification phase—the importance of *personalization* of the digital intervention typically comes forward. Saleem et al. (2021) conducted a scoping review to investigate different types of user engagement strategies. Their findings highlight that various personalization strategies—such as including adaptive feedback, tailored content, and user-specific reminders—can act as a central mechanism for sustained engagement in mental health interventions. Similarly, Braun et al. (2024) conducted a two-phase mixed-method study to obtain the needs of users for a mobile health application in which users first generated feature statements through structured brainstorming, followed by a second cohort that sorted and rated them. The resulting aggregated dissimilarity matrix was analyzed using multidimensional scaling and hierarchical clustering, yielding six conceptually distinct clusters. Feature clusters related to self-monitoring and personalization were rated as highly important—further underscoring their central role in driving user engagement and perceived app value.

To explore the full range of personalization variables implemented in mHealth interventions, Rivera-Romero et al. (2023) conducted a systematic review. From an initial pool of 1,139 articles, 62 studies were included in the narrative synthesis, all of which reported on mHealth solutions integrating at least one form of personalization. The results illustrated that demographic variables, user characteristics and health-related metrics each appeared in over half of the reviewed mHealth solutions, while behavioral, preference, contextual, and psychosocial items were less frequently employed.

To further clarify how these personal variables are actually implemented in eHealth solutions, Klooster et al. (2024) systematically reviewed 412 studies and uncovered 13 distinct clusters of personalization approaches. Their analysis revealed that user segmentation —for tailoring purposes—was overwhelmingly based on behavioral variables and individual determinants, while technology-related, user-interaction, environmental, and psychosocial variables appeared only sporadically (Figure 9). Furthermore, adaptation strategies were predominantly content-focused—primarily comparative feedback and personalized advice—while graphical enhancements and modifications to core functionality were observed in only a few cases.

Figure 9

An Overview of Frequently used Variables and Methods for Personalizing Digital Interventions



Note. Panel A illustrates that demographic variables, user characteristics and health-related metrics are most frequently used in mobile health applications. Panel B provides an overview of 13 cluster pairs, showing how different type of data are paired with specific system adaptation methods. The figures are adapted from Rivera-Romero et al. (2023) and Klooster et al. (2024), respectively.

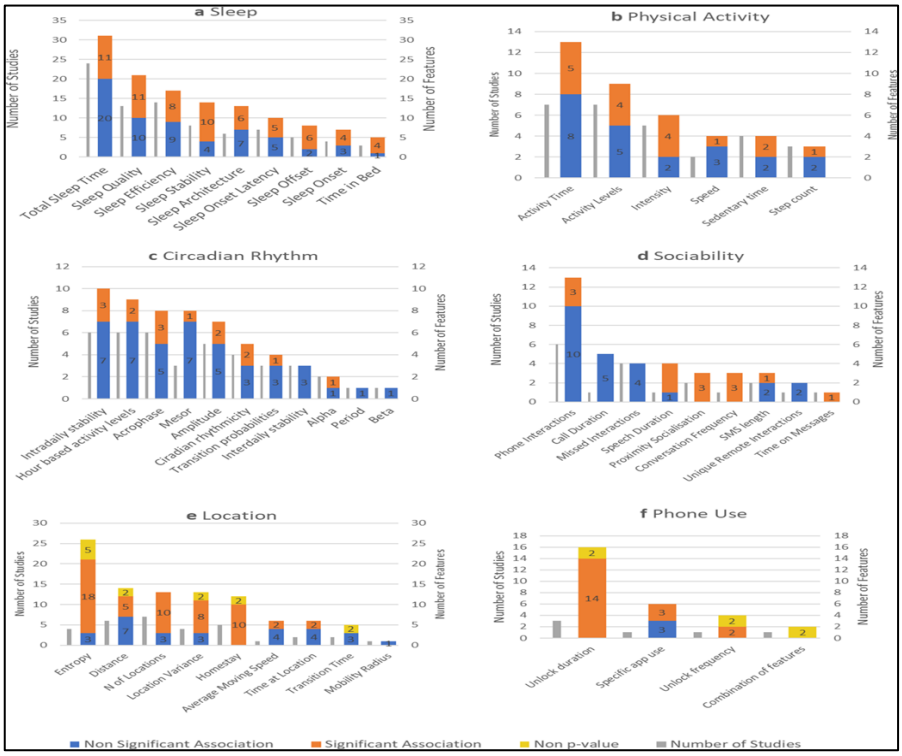
1.2.4.3 Improving Attrition Rates through Digital Passive Monitoring

Schroé et al. (2022) illustrated the importance of avoiding time-consuming questionnaires by observing that 55% of participants identified the repetitive and lengthy nature of surveys as their primary reason for discontinuing the intervention. For this reason, an interest in *passive monitoring* methods has been growing to personalize digital interventions (Sajnovic et al., 2024; Woll et al., 2025). Instead of relying on self-report, passive sensing uses the smartphones and wearable's built-in sensors to unobtrusively capture minute-by-minute data on everyday behaviors.

For example, De Angel et al. (2022) systematically reviewed 51 studies on depression employing only smartphone and wrist-worn sensors to extract low-level features—such as sleep stability, sleep offset, time in bed, intensity of physical activity, conversation frequency, socialization measured by average phone proximities, location entropy, homestay duration, and phone unlock duration—that were shown to significantly correlate with depressive symptom severity in the majority of reviewed studies (Figure 10). By aggregating these low-level streams into high-level behavioral markers, digital interventions can detect in-the-moment shifts in mood, enabling real-time personalization and mental support that continually adapts to the user's evolving state (Rogan et al.,2024; Rocchi et al., 2025).

Figure 10

Overview of Passive Monitoring Variables and their Association with Depression Severity



Note. This figure shows—for six behavior categories (sleep, activity, circadian rhythm, sociability, location, and phone use)—how often the included studies examined each feature, and classifies their link to depression severity as significant, non-significant, or assessed without p-values. Adapted from De Angel et al. (2022).

Rocchi et al. (2025) furthermore illustrated the usefulness of passive monitoring methods on a non-clinical population of young adults by deployed a Telegram chatbot that leveraged passive digital phenotyping—using statistical proxies for latent mental-health constructs (e.g., nocturnal phone-use patterns as an insomnia indicator). Four user clusters—distinguished by notification burden, messaging/social-media activity, sleep disruption, attention and energy levels, and symptom reports—informed a personalized exercise recommender system. Participants interacting with the chatbot experienced significant improvements in well-being and rated the tailored exercises as highly useful. Finally, although passively obtained low-level features are generally less informative for complex clinical symptoms—such as hallucinations and delusions—, Bladon et al. (2025) systematically nevertheless reviewed 203 studies passive monitoring in psychosis and schizophrenia.

Exceeding initial expectations, their review demonstrated that passive phenotyping offers clinical promise for continuous, real-time mental-health monitoring of these 'higher-order' clinical symptoms. They concluded their review by underscoring the imperative for standardized pre-processing protocols, rigorous quality-control measures, and enhanced transparency in reporting frameworks. Altogether, these findings highlight the potential of passive monitoring for mental health—via mobile and wearable technologies—to collect large amounts of longitudinal, single-subject, multivariate data enabling tailored digital interventions.

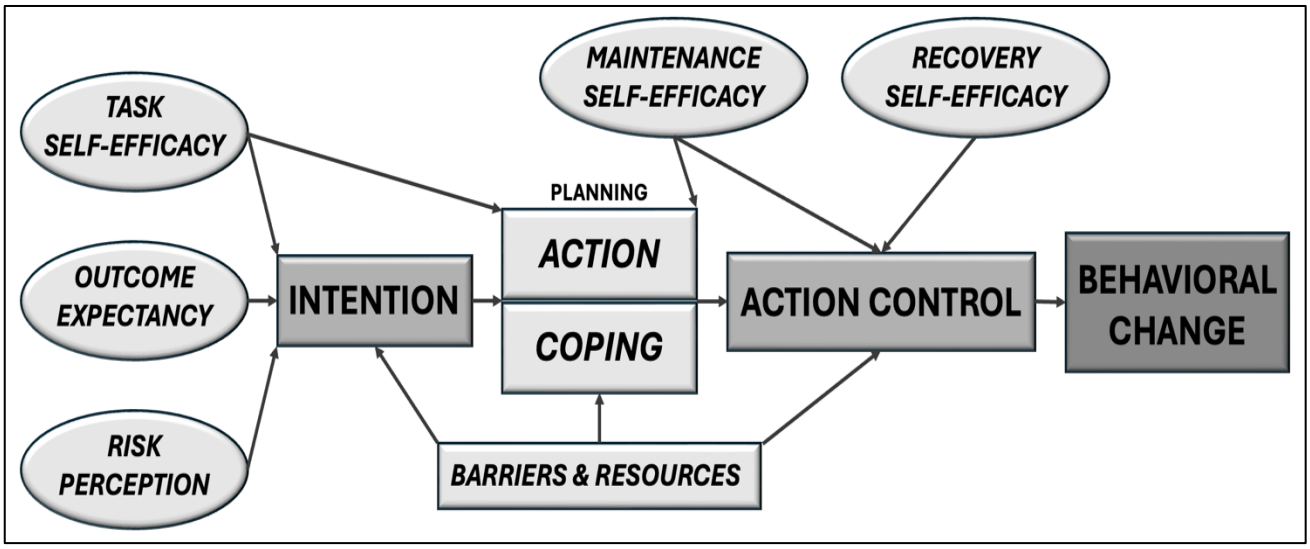
1.2.4.4 Aligning Digital Interventions with Theories of Behavioral Change

Translating these actionable insights—derived for idiographic purposes—into tailored digital interventions requires a robust theoretical foundation to effectively drive behavioral change. One such model—originally introduced by Prochaska and Diclemente (1983) on smoking behavior—is the Transtheoretical Model, which conceptualizes behavioral change as a dynamic, stage-based process: 1) *Precontemplation*, where there is no intention to change and often limited awareness of the problem; 2) *Contemplation*, marked by emerging awareness and ambivalence toward change; 3) *Preparation*, involving intention formation and initial steps; 4) *Action*, characterized by overt behavioral modification; and 5) *Maintenance*, which focuses on sustaining the new behavior and preventing relapse. By recognizing that individuals may progress through these stages in a non-linear and iterative manner, this perspective provides a flexible and adaptive framework for tailoring interventions to an individual's current state of readiness (Nicoara et al., 2024; Krishna et al., 2025).

A second theory on which many digital interventions for mental health-related problems are based—which has shown to empirically withstand rigorous testing—is the **Health Action Process Approach** (HAPA), initially proposed by Schwarzer et al. (2008). They explicitly distinguish between a motivational and a volitional phase (Rouvere et al., 2024). Within the volitional phase, action and coping planning are regarded as critical self-regulatory strategies that bridge the intention–behavior gap. Fernández et al. (2015) conducted a longitudinal study showing that both *action planning* and *coping planning* independently predicted change of behavior, with significant correlation estimates of $r = 0.47$, highlighting their practical relevance.

Figure 11

An Overview of the Health Action Process Approach — proposed by Schwarzer et al. (2008)



Note. The HAPA model distinguishes between a motivational phase—where intentions are formed—and a volitional phase—where intention is translated into action. The model emphasizes three forms of self-efficacy: action self-efficacy (pre-intentional), maintenance self-efficacy (during action), and recovery self-efficacy (after setbacks). A bipartite planning component models how users can prepare for undertaking action.

In both models, the importance of planning—specifically action planning and coping planning—emerges as a critical determinant of successful behavior change. Action planning involves concretely specifying when, where, and how to perform a desired behavior, while coping planning anticipates potential obstacles and formulates adaptive responses. These planning mechanisms serve to bridge the intention–behavior gap by increasing behavioral automatization and reducing cognitive load during execution. For this reason, semantic technologies intended to support digital interventions should be equipped with high-resolution representations of both cognitive-behavioral barriers and their corresponding coping strategies. Such ontological structures enable dynamic reasoning over personalized profiles, allowing systems to algorithmically match users with context-appropriate, theory-driven intervention components (Braun et al., 2024).

1.2.4.5 Evaluation of Large Language Model-based Digital Interventions

Finally, any digital health intervention that embeds large language models within its core architecture must be tested by a rigorous, multi-method evaluation framework that allows for subsequent iterative refinement. For instance, Haag et al. (2025) conducted a blinded study in which LLM-based *just-in-time adaptive intervention* (JITAI) messages were systematically compared against human-based JITAI's across a series of standardized use-case scenarios. The interventions were rated on seven-point Likert scales for four primary dimensions—appropriateness, engagement, projected clinical effectiveness, and professionalism. The LLM-based JITAI's were shown to significantly outperform human-based intervention suggestions on all four dimensions. In

contrast, an alternative approach for evaluation is to a adopt a comprehensive qualitative strategy, as demonstrated by Braun et al. (2025). In their study, they specified a concise set of detailed competency questions for each decision node of their system and then instantiated representative use cases to verify whether all competency requirements were sufficiently satisfied.

A standardized approach—both quantitative and qualitative—was introduced by Tam et al. (2024) to systematically evaluate LLM-based healthcare applications. Their **QUEST framework** is grounded in several core principles—a) Quality of information; b) Understanding and reasoning c) Expression style and persona d) Safety and harm; and e) Trust and confidence. Based on these dimensions, they proposed a three-phased workflow: 1) *Planning*, in which the team articulates the model's clinical goals and use-cases, selects a targeted set of QUEST dimensions, determines a statistically justified sample size, and specifies evaluator qualifications and training protocols; 2) *Implementation and adjudication*, during which the LLM is run on the sampled prompts, outputs are randomized and presented for blinded human rating, inter-rater agreement is quantified (e.g. Cohen's κ), and any low-agreement cases are iteratively discussed by expert adjudicators to reconcile discrepancies and refine the evaluation rubric; and 3) *Scoring and review*, where quantitative scores (e.g., means or medians across evaluators) and qualitative comments are aggregated, compared against automated benchmarks (e.g., F1), and synthesized into a detailed report that highlights both overall performance and specific failure modes, thereby guiding targeted, data-driven model improvements.

1.3 Research Questions

Given the exploratory nature of this technology-oriented master's thesis, the research questions—that try to fill in the identified gaps based on my literature study—are predominantly performance-oriented. For sake of clarity, all literature-based and methodological sub-questions—that will arise during the development of the first prototype of the optimization engine *'PHOENIX'*—fall under the following three overarching questions:

1. "How can **mental health**—and its corresponding corpus of **biopsychosocial predictors**—be properly formalized into a sufficiently comprehensive, interoperable, and non-redundant ontology for creating high-resolution, **single-subject representations** of (non-)clinical mental functioning?"

2. "How can **large language models** enhance—both during development and deployment—modern state-of-the-art systems that process real-time, longitudinal personal data (both structured and unstructured) that aim to **identify** personal treatment targets for mental health?"

3. "How can **large language models** enhance modern state-of-the-art systems that **translate** the identified treatment targets into a theoretically grounded personalized digital intervention that is displayed on a (read-only, non-interactive) screen of a mobile or web application?"

2. METHODS

2.1 Development of the Software's Analytical Component

To provide a clear answer to the three overarching research questions, a software solution will be developed that aims to tailor digital interventions in mental healthcare applications by leveraging large language models to deal with scalability problems during development and deployment. The analytical backbone of the software rests on three key components: 1) A fo*rmal representation of mental health*—and its corresponding corpus of biopsychosocial predictors (i.e., an ontology).; 2) A *hierarchical updating algorithm* that aims to quantify the momentary harmful impact of each individual variable on a set of criteria based on longitudinal multivariate time series data.; 3) A *modular agentic framework* that provides LLM's with a form of control to enhance their reasoning output during five key decision point of the software solution.

Combined, these three components form a data-driven optimization engine that operates by successively searching for personal treatment targets by exploring—together with the (data collecting) user—new variables to that show promise in resolving a given mental well-being problem. The engine is titled **'PHOENIX'**, which stands for *Personalized Hierarchical Optimization Engine for Navigating Insightful eXplorations* (Figure 1). Throughout the development of these analytical software components, the five pillars of the CeHRes roadmap , proposed by Kip et al. (2024), will be adhered to. Given the developmental nature of this first methodology section, only the most essential results will be reported for each of the three components.

2.1.1 Development of PHOENIX Ontology

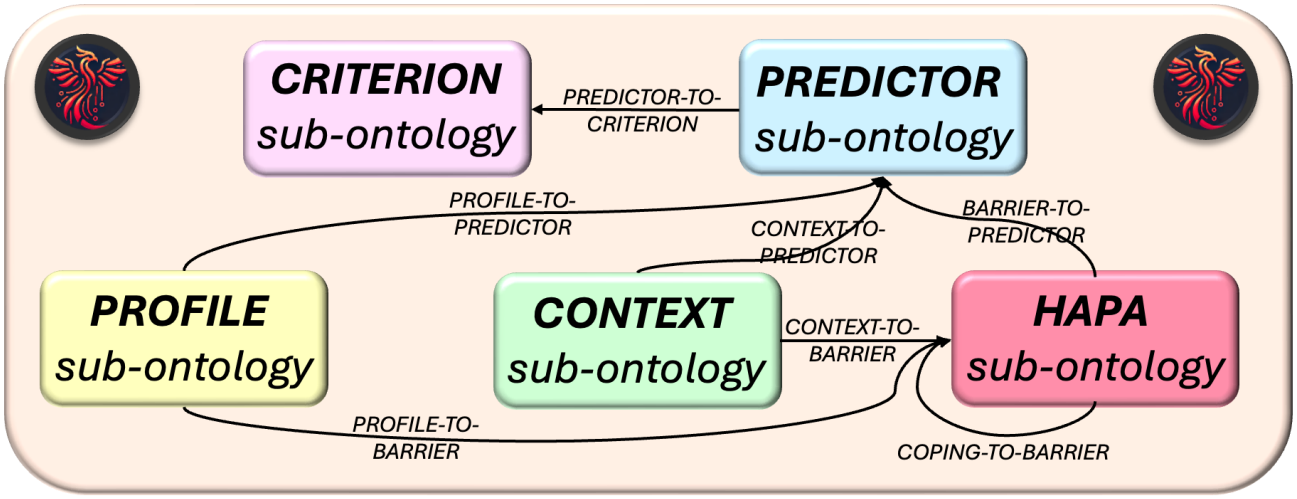
The desired outcome of this developmental process is to create a sufficiently comprehensive representation of mental health—and corresponding influential factors that could be of relevance for personalized context-aware recommendation systems. For this reason, five high-resolution interconnected taxonomies will be created—with support of large language models—with the intended purpose of enhancing the reasoning performance of five modular agentic frameworks in the custom-built optimization engine: 1) *CRITERION ontology*—aimed at representing mental states; 2) *PREDICTOR ontology*—aimed at representing potential influential treatment targets for complex mental state spaces for which no clear single treatment target exists; 3) *PERSON ontology*—aimed at representing general personal information such as demographics and preferences; 4) *CONTEXT ontology*—aimed at representing factors to contextualize recommendations; and 5) *HAPA ontology*—aimed at representing relevant barriers and coping options (Figure 12).

All of the abovementioned ontologies were implemented in OWL (i.e., Web Ontology Language)—a formal knowledge representation language designed for expressing rich, logical structures within a domain. OWL allows for the definition of classes, properties, individuals, and axioms—including subclass hierarchies, domain/range constraints, and logical restrictions such as disjointness and cardinality. This level of semantic expressiveness enables automated reasoning and consistency checking across complex conceptual schemas.

To support the LLM-based scalability during graph construction, the OWL ontologies were programmatically constructed using the Python library '*rdflib'*, which allows for the creation and manipulation of RDF (i.e., Resource Description Framework) objects. Each ontology was serialized into an '.owl' file format, which can be directly visualized and manually edited using ontology development tools such as *Protégé*, thereby enabling both human inspection and machine-driven inference (Gennari et al., 2003).

Figure 12

A High-Level Overview of the PHOENIX Ontology with the Intra-Ontological Mappings



Note. The PHOENIX ontology consists of five interconnected sub-ontologies—with the objective to be employed inside an agentic LLM-based framework to identify impactful predictors for a given mental state; and subsequently translate those predictors (i.e., treatments targets) into a personalized and context-aware 'digital message'-as-intervention. All seven necessary mappings are indicated by the directed lines.

2.1.1.1 'Criterion' Sub-Ontology — Mental Health

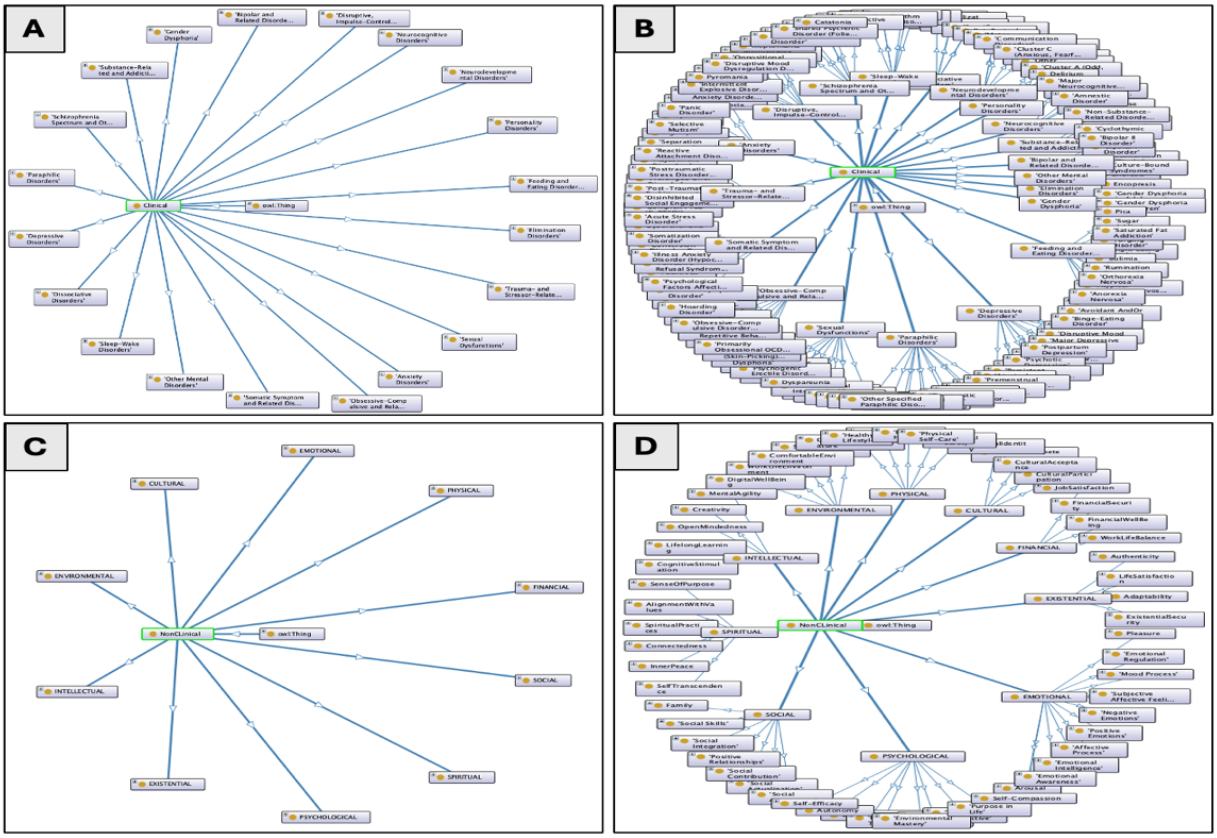
The process of creating a formal representation of mental health started by constructing a high-level hierarchical taxonomy of *DSM-V-TR's clinical categories* (American Psychological Association, 2023). In total, 20 primary domains were extracted with a total of 304 mental disorders. Subsequently, the encoded knowledge of the large language model 'gpt-4o' (OpenAI, 2025) was employed to generate a high-resolution representation of any given mental disorder. A pilot test was conducted on an initial set of 50 randomly selected mental disorders to test whether all key symptoms were present in the LLM-generated list of symptoms. As predicted, based on previous research on their accurate clinical knowledge (Van Veen et al., 2024; Workum et al., 2025), the LLM had an excellent performance where 99.38% of all the actual symptoms were successfully generated. For this reason, the large language model was then instructed to generate a set of six to fifteen clinical symptoms that collectively represent the full expression for all 304 prompted disorders.

After this, given that the World Health Organization also includes a *non-clinical component* in their definition of mental health, a high-level taxonomy of 10 non-clinical domains were manually crafted. These domains

were then also separately passed through the large language model that was instructed to create—in a recursive fashion (until satisfied)—a mid to low-level hierarchical taxonomy. By combining the leaf nodes (i.e., classes with no subclasses, representing the most specific concepts in an ontology) from the non-clinical and clinical taxonomy, a list of 5275 items was created (Figure 13).

Figure 13

An Overview of the Unprocessed Taxonomy of Mental Health —Populated by a Large Language Model



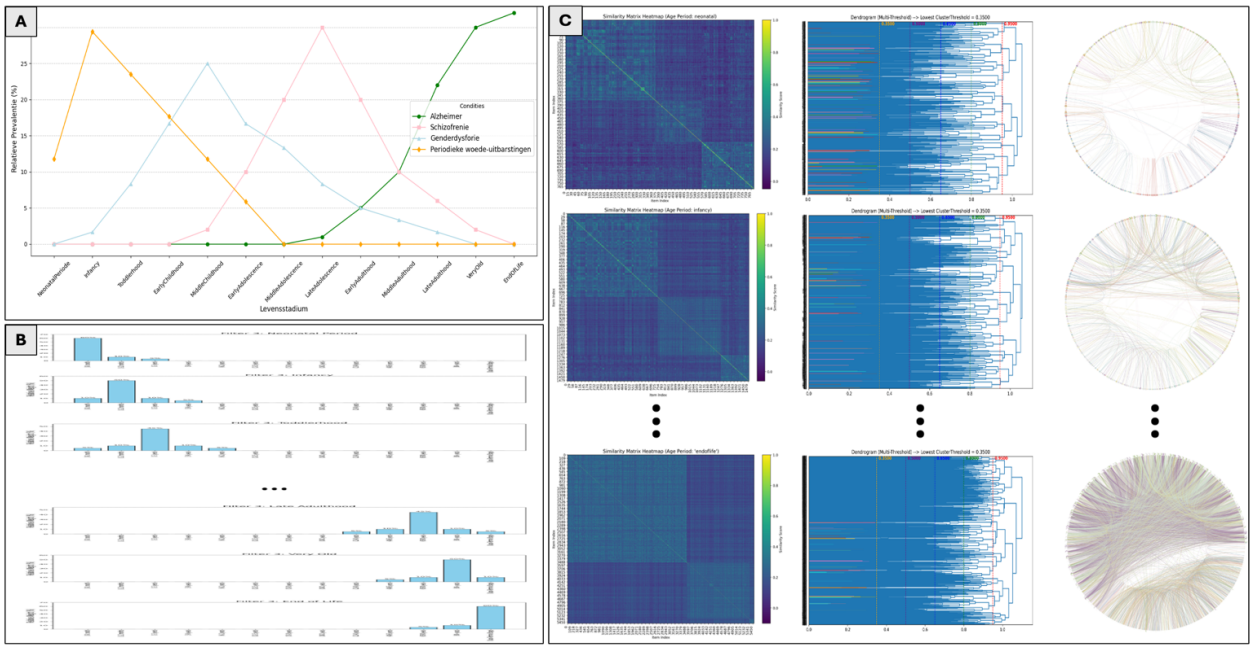
Note. Panel A and C provide a high-level overview of the formalized representation of clinical and non-clinical mental states, respectively. In panel B and D, the second layer of the taxonomies are visualized where all primary nodes of the 20 clinical domains and 10 non-clinical domains were expanded, respectively.

After selecting the most relevant items for all age groups, 13 age-specific sub-ontologies were created. These were then, once again, populated by two additional streams of LLM-generated information: 1) An LLM was instructed to generate mid to low-level non-clinical taxonomies for four dimensions—cognitive, social, emotional and physical—based on Erik Erikson's Psychosocial Theory of Development (Erikson, 1963; Orenstein & Lewis, 2022) for each of the 13 age groups; 2) An LLM was instructed to generate a single mid to low-level taxonomy of idiosyncratic non-clinical values for each age group. This procedure of LLM-based content population aims to increase the resolution with which mental health can be conceptualized. The result was an average number of 1152.84 (*SD*=253.13) items in the 13 age-specific sub-ontologies.

As demonstrated by Forbes et al. (2023), the DSM-V-TR contains a substantial amount of between-disorder overlap of symptoms. This redundancy was also noticeable in the 13 age-specific sub-ontologies—even when comparing items between the non-clinical taxonomy and clinical taxonomy (e.g., restlessness, worry, and social withdrawal). To remove this *semantic redundancy* in the ontologies, all labels of their leaf nodes were embedded by OpenAI's embedding model 'text-embedding-3-large' (2025) to create a high-dimensional representation of their semantics. For each pairwise combination, their cosine similarity coefficient was computed—which is frequently used in linguistic research to obtain an overview of how dissimilar two high-dimensional representations are (Elekes et al., 2017; Steck et al., 2025). Then, an agglomerative complete-linkage clustering algorithm (Grosswendt & Roeglin, 2017) was applied to each of the resulting dissimilarity matrices—which iteratively merges the most similar class pairs into a single hierarchy of nested clusters.

Figure 14

An Overview of the Procedure to Construct Non-Redundant Age-Specific Ontologies



Note. Panel A illustrates four examples—for Alzheimer's disease, schizophrenia, gender dysphoria, and disruptive mood dysregulation disorder—of the generated frequency distributions based on relative prevalence estimates for each of the 13 age groups. Panel B shows the second type of frequency distributions where for older age groups the relative importance score starts shifting to the right. These two distribution types are convolved with each other to match specific items to specific age groups. Panel C demonstrates the procedure that was used to cluster semantically overlapping items in all of the 13 age-specific sub-ontologies. The three columns contain 1) the embedding-based dissimilarity matrix, 2) the dendrogram outputs of the clustering algorithm and 3) the circle plot where each connection indicates a redundancy (i.e., inside same cluster where θ=0.35); from the youngest (bottom row) to oldest (upper row) age group.

The most optimal intra-cluster threshold was determined to be $\theta = 0.35$—by systematically comparing the generated list at each leaf node cluster—for grouping redundant pairs of class entities (Figure 14). Finally, to deal with clusters that still contained two semantically distinct concepts, each cluster was passed through an LLM that was instructed to split any remaining types of redundancy. Wherein, if any conceptual non-overlap was identified (e.g., insomnia vs hypersomnia), those respective items were clustered into different lists. This resulted in a set of 13 age-specific ontologies that were assumed to be covering the majority of possible mental states. Clinical classes from alternative frameworks to conceptualize mental health—such as from Research Domain Criteria (Cuthbert et al., 2020; Morris et al., 2022)—are currently not used incorporated.

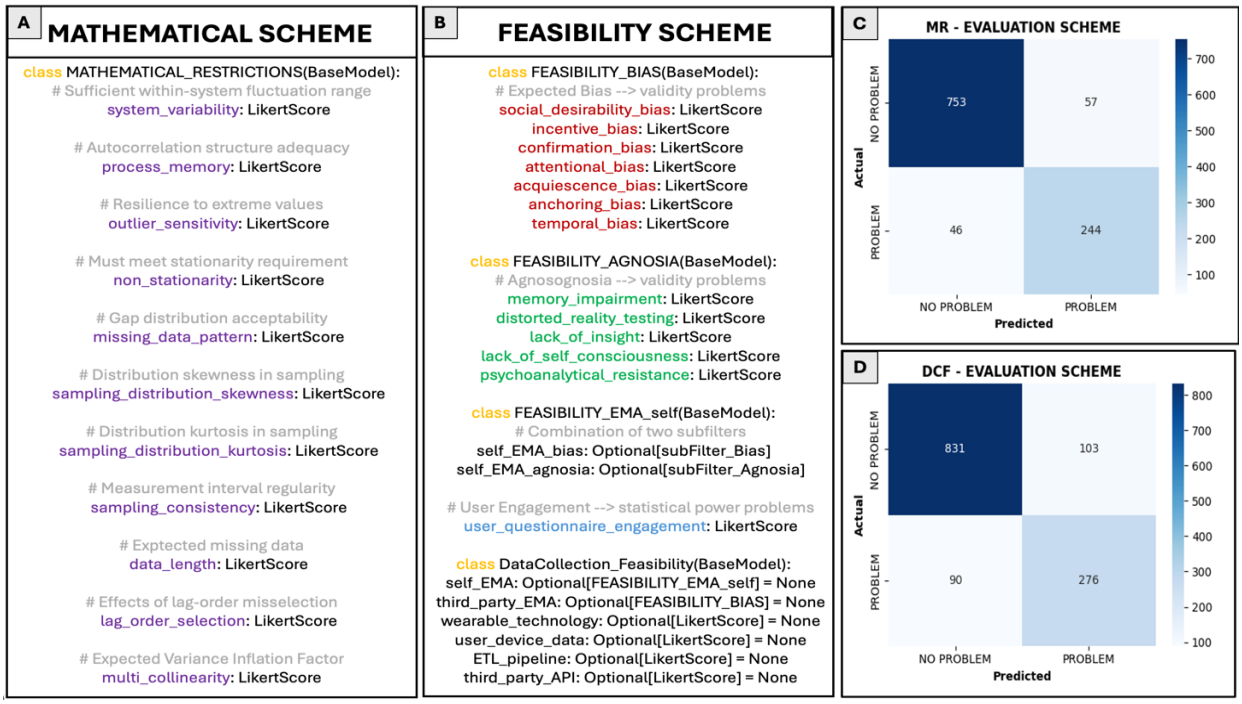
An important objective of this Criterion ontology is to be used by an LLM-based agent that operationalizes a given textual (i.e., of data type 'string') formulation of a mental state into a set of criterions. A subset of these variables will subsequently be selected to create the criterion part of the initial observation model—which will be used to determine what variables are of importance to model the momentary mental state of an individual. Any type of data collection procedure can be used to obtain the multivariate time series. The following non-exhaustive list highlights a few of those data collection methods: a) ecological momentary assessment; b) wearable technology; c) (semi-)automated pipelines that transform unstructured data into a single estimate for each variable for each time point; and d) third-party API's to transfer structured data from external databases.

For this reason, —that collecting multivariate time series is a key component of the optimization engine—two schemes of evaluation are applied to identify what variables in the Criterion ontology are 1) suitable for the multivariate time series analysis (i.e., time-varying gVAR), and 2) have a high feasibility for collecting enough data with sufficient validity. The *mathematical restriction* (MR) *scheme* is used to rank all leaf nodes according to how much problems would arise during the time series analysis—either by violation of assumptions or other expected sampling patterns that could disrupt the analysis. The *data collection feasibility* (DCF) *scheme* is used to rank all leaf nodes according to how much problems would arise during six types of data collection methods that could limit the quality—predominantly internal validity—of the obtained data.

Given that the amount of leaf nodes to be evaluated in the 13 age-specific ontologies is too much to manually evaluate (M=824.92; SD=214.33), 'gpt-4o' was used. A pilot test was conducted to estimate the evaluation abilities of the LLM with a subset of 100 randomly selected leaf nodes. The LLM was instructed to generate, for each leaf node, 11 and 13 evaluation scores for the MR scheme and DCF scheme, respectively; resulting in a total of 2400 evaluation scores ($N = 100 \text{ items} * (11 \text{ MR\_criteria} + 13 \text{ DCF\_criteria})$). Each evaluation score—employing a 9-point Likert scale—indicated the estimated likelihood that, for any given leaf node, a problem could arise. The 100 selected leaf nodes were also annotated with a binary label indicating whether the node was, for each point of evaluation, likely to pose a problem. Binary confusion matrices were constructed for each evaluation scheme—with an average classification accuracy of 90.64% for the MR scheme, and 85.15% for the DCF scheme (Figure 15).

Figure 15

An Overview of the Two Evaluation Schemes and Pilot Test to Identify Suitable Criterions



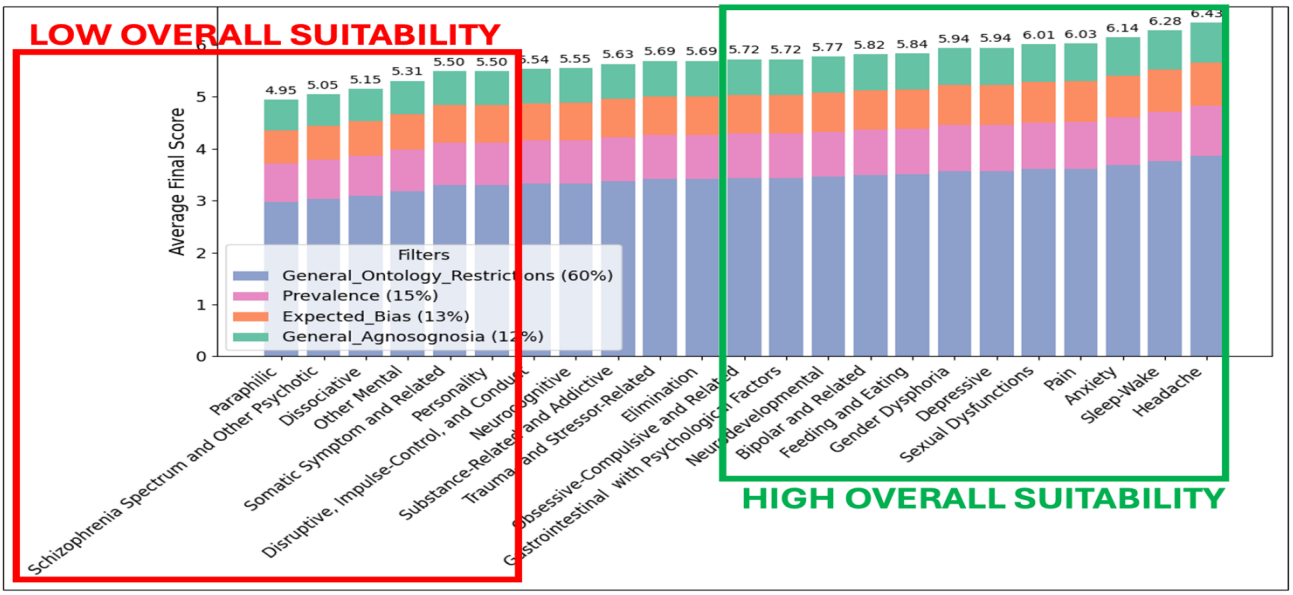
Note. Panel A and B are copies of the Pydantic models that were used for schema validation during the LLM-based generation of evaluation scores. For each evaluation dimension, the LLM was instructed to estimate whether a problem could arise on a 9-point Likert scale; Panel C and D show the results of a pilot test where 100 randomly selected leaf_node ID's were manually annotated in a binary fashion to evaluate the LLM performance. The LLM-based estimates were binarized (i.e., 'PROBLEM' if *Likert_score* > 5; 'NO PROBLEM' if *Likert_score* < 5) to generate two binary confusion matrices. One for the 11 dimensions of the MR scheme (panel C), and one for the 13 evaluation dimensions of the DCF scheme (panel D).

Given that the LLM performance in the pilot test was sufficiently high, the two evaluation schemes were applied to all the remaining leaf nodes with the same prompting procedure as in the pilot test. To rank all classes according to their overall suitability, each of the 24 evaluation scores—for each leaf node—were multiplied by a fixed distribution of importance weights that allows certain evaluation dimensions to be prioritized (e.g., sufficient system variability or social desirability bias). An additional estimate of overall prevalence was generated for each leaf node for practical purposes—to ensure that the development is directed towards larger (non-)clinical populations—such that leaf nodes with a higher clinical absolute prevalence would get a higher ranking score. In total, 3421 unique labels (i.e., classes) were evaluated; resulting in a total amount of approximately 85.000 scores (*N = 3421 label * (24 evaluation scores + 1 absolute prevalence score)*) were generated. Weighted averages were computed for each unique label and subsequently sorted to obtain a ranked list of least to most suitable—for the intended purposes of the ontology— class entities.

Classes that were expected to be more feasible in general, when looking at the evaluation dimensions, were indeed ranked in a higher place. When retrogradely—i.e., based on what primary domain the leaf node originated from—tracing what clinical disorders are most suited for the multivariate time series analyses (incl. procedure of obtaining valid data), results aligned with expectations. Mood disorders—such as anxiety or depression disorders—for which no strong feasibility issues were expected to arise, received higher ranking scores than mental disorders which are either 1) highly static (i.e., low intra-weekly system variation) such as 'personality disorders' or 2) impractical when looking at the individual's ability to participate in observational studies for extended periods of time where time series data must be obtained with sufficient internal validity such as 'paraphilic disorders', or 'disruptive, impulse-control and conduct disorders' (Figure 16).

Figure 16

Results of Ranking Scores to Identify what Clinical Domains are Most Suitable for the Software Solution



Note. Disorder domains with static mental states for which feasibility issues could arise during data collection resulted in noticeable lower overall suitability ranking scores (e.g., personality disorders, paraphilic disorders, schizophrenia-related/ psychotic disorders). The most suitable classes for the intended purpose of the Criterion ontology were mood disorders—such as anxiety and depression disorders—and headache disorders. All the importance weights for the 25 (*24+1*) evaluation dimensions were summed into four groups: 1) 'mathematical restriction' importance—60%; 2) 'prevalence' importance—15%; 3) 'expected bias' importance—13%; and 4) 'anosognosia' (i.e., a mental state in which an individual is not able to accurately report their state due to lack of insight, consciousness, or perceptual abilities) importance—12%.

Additionally, three ICD-11 domains (World Health Organization, 2024) were also included—with the same pre-processing steps—which were expected to receive similar ranking scores as the mood disorders: a) gastro-intestinal disorders; b) pain disorders c); and headache disorders. The main reason for their inclusion is that a substantial proportion of their clinical categories are typically operationalized by non-static variables such as

‘intensity’ or ‘frequency’—of which their severity is known to arise from sophisticated (often idiopathic) etiological pathways (Karcioglu et al., 2018; Holland et al., 2021; Robinson et al., 2024). Surprisingly, when including all three domains, with a total of 64 additional clinical diseases, headache disorders—which were expected to 1) show sufficient intra-weekly system fluctuation in severity, and 2) be relatively accurately assessed—obtained the highest overall suitability score; even higher than mood disorders (Figure 16).

2.1.1.2 ‘Predictor’ Sub-Ontology — Biopsychosocial Treatment Targets

The developmental process of the Predictor ontology will be similar to that of the Criterion ontology. A high-level taxonomy will be constructed based on Engel’s **Biopsychosocial Model**—aimed to start building from an empirically robust structure of interdisciplinary therapeutic approaches (Engels, 1977; Alvarez et al., 2012; Fanali et al., 2024; Fava et al., 2024; Pullman et al., 2025). This model conceptualizes mental health as the result of dynamic interactions between biological, psychological, and social factors. Within this high-level taxonomy, the *biological* layer includes mechanisms that modulate neural and physiological functioning—such as neuromodulation, physical activity, sleep-wake regulation, and nutritional status. The *psychological* layer targets intra-personal processes—such as cognitive reframing, behavioral activation, emotional regulation, and self-monitoring. Lastly, the *social* layer emphasizes contextual and relational factors—such as inter-personal support, peer connection, social skill development, and environmental adaptation.

Starting from this high-level taxonomy, the LLM ‘gpt-4o’ will populate—in a recursive manner until satisfied with the depth (i.e., resolution)—this structure by complementing it with a low to mid-level taxonomy. After this procedure, each leaf node of the obtained Criterion ontology will be passed through the same LLM with a similar instruction to generate a low to mid-level taxonomy—but now highly specific to the prompted criterion. This will result in a single ontology that may have ontology paths with no clear applicability to certain age groups. Therefore, 13 age-specific sub-ontologies will be created by employing the same LLM-based splitting procedure as before. A pilot test will be conducted beforehand to evaluate the performance of the LLM to estimate age-specific relevance scores. These generated distributions, which will come from 100 randomly selected predictor classes, will be compared with a manually annotated dataset. A two-sample Kolmogorov-Smirnov test will be used for determining whether there is a significant difference between the annotated vs generated distributions; and the total deviation will be quantified using the Kullback-Leibler divergence (Broniatowski, 2006). Finally, given that the resulting structure is expected to contain structural overlap, the same redundancy removal procedure will be applied as demonstrated in the previous section.

An important objective of this Predictor ontology is to be used by an LLM-based agent to identify what potential predictors have the highest negative impact on a given mental state. Therefore, since the data collection of multivariate timeseries is an inherent part to this identification procedure, the two evaluation schemes—mathematical restriction scheme and data collection feasibility scheme—will be applied to the 13 age-specific non-redundant sub-ontologies. An additional evaluation scheme will be applied to obtain an

estimation of the potential, for a given predictor, to become a treatment target that lies within a proximal zone of responsibility. This third *treatment translation* (TT) *evaluation scheme* is an important component to restrict the scope of this sub-ontology for enhancing the applicability of the PHOENIX ontology—where, by design, predictor variables should be used during observation of which at least some of them can be directly translated into treatment targets. Otherwise, the results of the multivariate time series analysis, irrespective of how the data was obtained, would not provide the client or healthcare provider with an actionable insight.

Once these three evaluation schemes are applied to the Predictor ontology, the next step is to ensure that the remaining set of biopsychosocial predictors fulfills the requirement of having a consistent abstract-concrete gradient—throughout the full Predictor ontology. That is, since the hierarchical updating algorithm aims to successively update the observation models from abstract to concreter versions (e.g., from 'nutritional status' to a sub-predictor 'calorie intake'; which is part of a concreter subset of 'nutritional status'), the Predictor ontology must meet the requirement of having a hierarchical structure where nodes in higher layers are more abstract than nodes in lower layers. For this reason, a *vertical migration algorithm* will be developed that uses an LLM to check this assumption for any prompted subset from the ontology—where for any local violation of the concrete-abstract gradient assumption, the LLM migrates the necessary predictors to either their parent layer or child layer. The first validation check begins at the first layer and iteratively progresses through each subsequent layer, continuing this cycle until convergence is achieved based on the global violation error—defined as the total number of migrated predictors across a full iteration. A pilot test will be conducted that evaluates the performance of the LLM on 100 randomly chosen subsets to 1) detect local violations in concrete-abstract gradient direction, and 2) correctly migrate predictors if any local violation is detected.

Finally, the resulting predictors must be mapped onto the criterions for each age-specific ontology—the first mapping procedure of the PHOENIX ontology *Predictor-to-Criterion mapping*. For this reason, the LLM 'gpt-4o' will be instructed to generate an estimate of the overall causal relevance of a certain predictor for a given criterion using a high-resolution 100-point binary semantic scale. This extensive pairwise procedure will result in 13 age-specific $N_c \, X \, N_p$ matrices with the objective to be used by the AI-agent that must find a relevant set of predictors for any given set of criterions during the construction of an observation model. A pilot test, using 100 randomly selected criterion-predictor—for each of 13 age-specific ontologies—will determine the performance of the LLM to estimate causal relevance score. An annotated dataset will be created to compare performance by computing the root mean squared error (RMSE).

2.1.1.3 'Profile' Sub-Ontology — General User Profile

To create a formal representation of the expected user profile, a similar LLM-based construction procedure will be used as for the criterion and Predictor ontology. A high-level taxonomy will be assembled based a systematic review by Rivera-Romera et al. (2023) on what personal variables are frequently used for digital intervention. Six primary domains will be used to create the high-level structure: a) *demographic data* such as

age and gender; b) *user characteristics* such as height; c) *health-related data* such as treatment history; d) *user preferences* such as preference for frequent feedback moments; e) *behavior change factors* such as motivation levels; and f) *intervention-related data* engagement with eHealth application (Figure 9). These domains will be passed through the LLM 'gpt-4o' that will be instructed to generate a low to mid-level taxonomy of general user profile data, given a description of the objective and architecture of the optimization engine PHOENIX.

The resulting Profile ontology will undergo two dedicated mapping procedures: 1) *Profile-to-Predictor mapping*—This mapping links user profile elements to the Predictor ontology to support the LLM-based agentic framework in tracking personalized predictor relevance over time. Retaining a broad range of socio-historical user information has been shown to significantly improve user satisfaction when interacting with conversation-oriented AI-agents (Zheng et al., 2024; Jiang et al., 2024). 2) *Profile-to-Barrier mapping*—This second mapping aligns user profiles with potential barriers, allowing the agentic system to more effectively identify obstacles and tailor intervention strategies. This mapping plays an important role for effectively personalizing digital interventions for mental health. For both mapping procedures, an initial pilot test will be conducted to assess the mapping performance of the LLM. For this, an annotated dataset—based on 100 randomly selected profile-predictor pairs—will be created for each mapping type to compare the performance which will once again be computed by the RMSE metric.

2.1.1.4 'Context' Sub-Ontology — Internal and External Environment

To formally represent contextual influences relevant to digital interventions in mental health, a high-level taxonomy encompassing 20 primary domains will be defined. This taxonomy will be designed to capture a substantial portion of the full scope of relevant contextual factors—such as time, location, and weather. This high-level structure will be enriched using the same LLM-based procedure applied in the previous ontology components, whereby 'gpt-4o' will be prompted to generate a low- to mid-level taxonomy for each primary domain. The resulting Context ontology will also undergo two similar mapping procedures: 1) a Context-to-Predictor *mapping*—which links contextual features to the Predictor ontology to help the LLM-based agent interpret dynamic relevance in light of environmental dynamics, and (2) a *Context-to-Barrier mapping*—which connects contextual elements to potential behavioral or structural barriers that may impede intervention delivery or adherence. Both mappings are intended to support the contextualization of digital interventions. For both mapping procedures, an initial pilot test will be conducted to assess the mapping performance of the LLM. For this, an annotated dataset—based on 100 randomly selected profile-predictor pairs—will be created for each mapping type to compare the performance which will once again be computed by the RMSE metric.

2.1.1.5 'HAPA-based' Sub-Ontology — Theory-grounded Factors for Behavioral Change

Once the most impactful predictors—referring to the potential treatment targets—based on both unstructured data (e.g., textual information) and structured data (e.g., *negative impact coefficients* based on multivariate time series analysis)—these must be translated into a tailored digital intervention. To theoretically ground this

process, the HAPA ontology will support the LLM-based agent in generating messages by providing a formal representation of relevant barriers and corresponding coping strategies associated with each treatment target Figure 11). Specifically, two ontology mapping procedures will be implemented: (1) *Barrier-to-Predictor mapping*, with the objective to link each treatment target to its most commonly associated psychological or behavioral barriers, and (2) a final *Coping-to-Barrier mapping*, which connects identified barriers to evidence-based coping strategies that can be embedded in the intervention message. For both of the mapping procedure, once again, an initial pilot test will estimate the LLM performance.

2.1.2 Development of Hierarchical Updating Algorithm

2.1.2.1 Multivariate Time Series Analysis

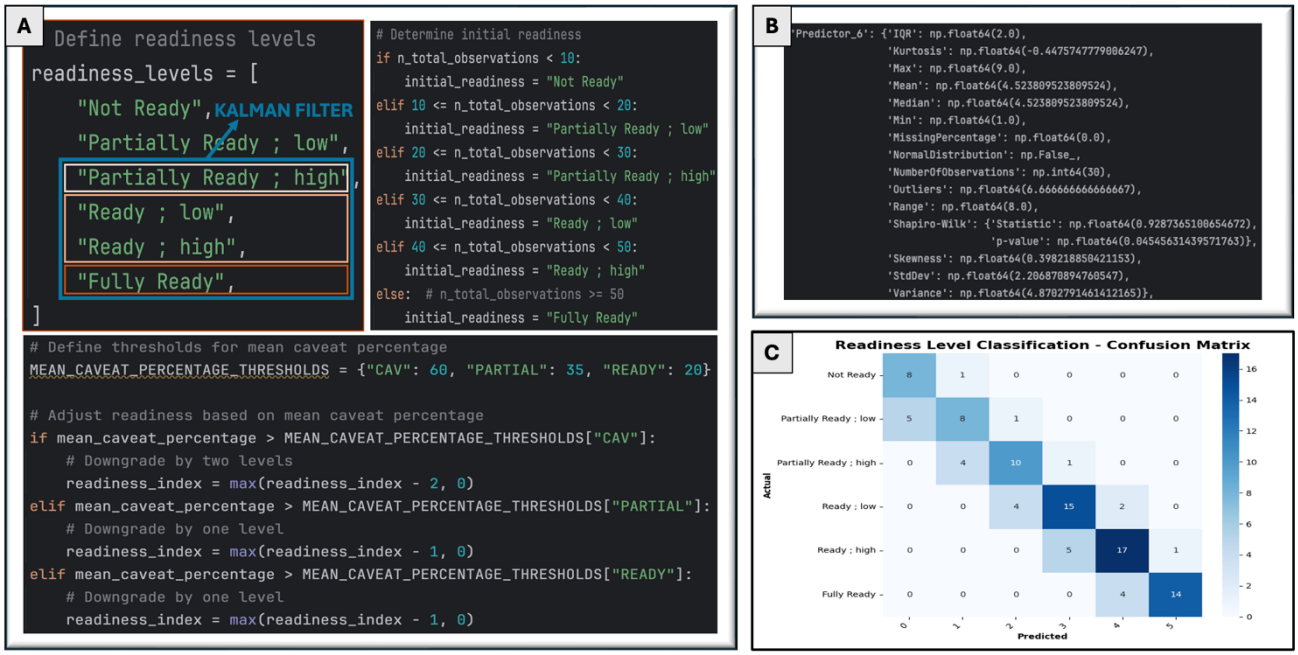
Since the hierarchical optimization engine will split up the intended full observation period into multiple observation periods—so that the model content can be updated successively—, there is a strong likelihood that the statistical requirements to perform the time-varying network analysis may not be fully satisfied. For this reason, any obtained data will be classified into six *levels of readiness*. Depending on the initial amount of observations and average caveat levels—computed by multiple statistical factors—, four possibilities may occur: 1) If 'NOT READY' or 'PARTIALLY READY – LOW', then no analysis will be conducted where the individual will be requested to obtain more datapoints; 2) If 'PARTIALLY READY – HIGH', only a regular correlation matrix will be computed to estimate the strength of relations between the variables; 3) If 'READY – LOW' or 'READY – HIGH', then a vector autoregression model will be fitted to the full dataset using the R R package *graphicalVAR*; 4) If 'READY – HIGH' or 'FULLY READY', then kernel-smoothed, locally weighted regressions will be performed to estimate a series of gVAR coefficient and precision matrices at several equidistant estimation points using the R package *mgm* (Haslbeck et al., 2020)

Missing data will be imputed with a *Kalman filter*, which has been demonstrated to outperform other imputation methods in settings with lower levels of statistical power (Agbailu, 2021). A pilot test was conducted to examine the classification performance of this procedure by randomly generating 100 multivariate time series—with a fixed set of four variables—that could vary in 1) the maximal length of observations; ranging from 5 to 100, and 2) the number of missing rows (i.e., missing univariate observation points); ranging from 0% to 50%. Based on 15 metrics for each of the four univariate time series, all the 100 multivariate time series were manually annotated with one of the six abovementioned labels to evaluate the performance of the procedure; results were plotted using a confusion matrix (Figure 17). A mean accuracy of 72% was found, which is concluded to be sufficient given that there were six classes, and all misclassification occurred in either its lower or upper neighborhood level. Noticeably, the empirical likelihood of being classified as less ready than the actual readiness level was 3.67 times higher—computed by dividing the number of misclassified examples that were predicted as less ready (22) by the remaining number of misclassified examples being classified as more ready (6). This is interpreted as the readiness estimation

procedure being relatively conservative—which is, from a medical point of view, a beneficial trait wherein low-powered analysis findings will be less likely to be used for interpretation.

Figure 17

An Overview of the Automated Readiness Estimation Procedure and the Classification Performance



Note. Panel A demonstrates the procedure where six levels are initialized for classification purposes to obtain an estimate of whether a multivariate time is ready to be analyzed, and to what extent (i.e., what type of analysis would be optimal). Then based on several metrics (shown in panel B with an example variable) and their corresponding thresholds for caveat classification, an overall caveat estimation can be computed for each univariate signal. Together with an initial readiness level based on the total number of observations—drawing on literature about minimal sample size targets (Christensen et al., 2024)—, the mean caveat percentage of each univariate signal results in an estimation of readiness. Panel C contains the confusion matrix from the pilot test where 100 multivariate time series were manually annotated to evaluate its performance.

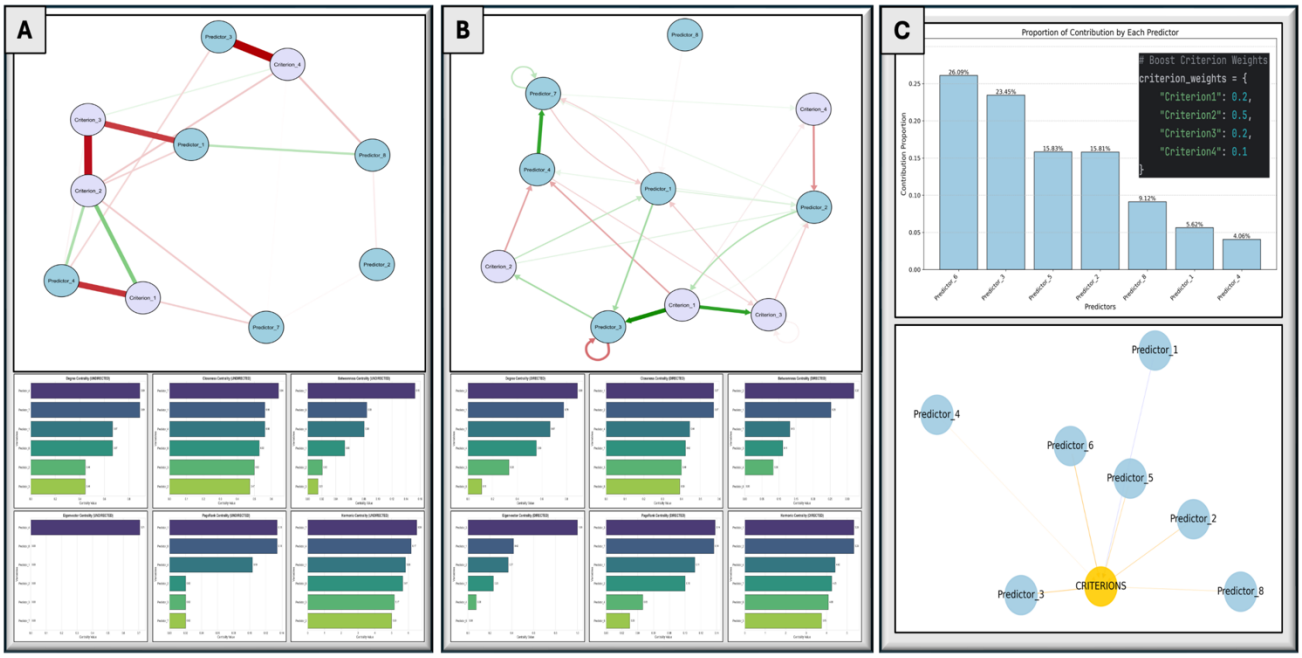
2.1.2.2 Quantification of Momentary Impact

Once the estimation matrices are obtained, then the overall *momentary impact* (MI) coefficients can be computed. These MI coefficients will support the LLM to identify treatment targets, where each predictor is assigned with a single score that summarizes the results of the full multivariate time series analysis. This coefficient is a combination of 1) six weighted (and softmaxed) centrality coefficients, and 2) a single mean value of the absolute relational strength to every criterion (Figure 18). For the temporal matrix, only the estimated strengths will be used where the predictor Granger causes the criterion. If the obtained multivariate data was sufficiently ready for a vector autoregression analysis—for which two matrices are computed (i.e., temporal and contemporaneous)—, then their single estimation score is given an equal weight of 0.5. If the obtained multivariate data was sufficiently ready for a time-varying analysis, then the single estimation

coefficient for each local analysis will be linearly proportional to the temporal distance from the mean local observation time and last observation time of the full data-analysis period; higher weights for more recent observations. Finally, the method for estimating the MI coefficients also incorporates the possibility to give higher *boosting weights* to certain criterions to ensure that some of them are prioritized over others. This can be medically advantageous in scenario's where multiple criterions are included in the observation model for which only some of them are relevant—for example where all symptoms of depression are being modelled, yet only a subset of them are, for one individual, below clinical thresholds—and therefore more relevant.

Figure 18

An (partial) Overview of Estimating the Momentary Impact Coefficients



Note. Panel A and B shows the results of the six computed centrality metrics with pseudo-data based on a contemporaneous and temporal matrix, respectively—Degree, Closeness, Betweenness, Eigenvector, PageRank and Harmonic centrality. Panel C contains an example of a Python dictionary with four criterion boosting weights, complemented by exemplary results of an array of MI coefficients that indicates how important each predictor was based on the full analysis of a single-subject's multivariate time series.

The resulting analysis scripts—which are a combination of Python and R code—to quantify what predictors have the strongest impact on a set of criteria that reflects an individual's mental state, will be made publicly available on GitHub under the GPL-3.0 license (Free Software Foundation, 2007):

> *"The GNU **General Public License** is a free, copyleft license for software and other kinds of works. The licenses for most software and other practical works are designed to take away your freedom to share and change the works. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change all versions of a program to make sure it remains free software for all its users."*

2.1.3 Development of Modular Agentic Framework

2.1.3.1 Operationalization of (Non-)Clinical Mental Health Problem

WERK IK DEZE ZOMER UIT

PARAGRAPH – DESCRIPTION OF FILTERING PROCEDURE

PARAGRAPH – DESCRIPTION OF GRAPH-RAG PROCEDURE

PARAGRAPH – DESCRIPTION OF VALIDATION SCHEME

WORDT VERDER UITGEBREID

2.1.3.2 Construction of Initial Observational Model

WERK IK DEZE ZOMER UIT

PARAGRAPH – DESCRIPTION MAPPING-BASED FILTER

PARAGRAPH – DESCRIPTION OF CHAIN-OF-THOUGHT + GRAPH-RAG PROCEDURE

WORDT VERDER UITGEBREID

2.1.3.3 Identification of Personalized Treatment Targets

WERK IK DEZE ZOMER UIT

PARAGRAPH – DESCRIPTION OF CHAIN-OF-THOUGHT + GRAPH-RAG PROCEDURE

WORDT VERDER UITGEBREID

2.1.3.4 Construction of Updated Observational Model

WERK IK DEZE ZOMER UIT

PARAGRAPH – DESCRIPTION INCORPORATING MULTIVARIATE ANALYSIS RESULTS + ETL PIPELINE TO CREATE UPDATED HIGH-RESOLUTION OVERVIEW OF MENTAL STATE

PARAGRAPH – DESCRIPTION OF CHAIN-OF-THOUGHT + GRAPH-RAG PROCEDURE

WORDT VERDER UITGEBREID

2.1.3.5 Translation of Identified Targets into Tailored Digital Intervention

WERK IK DEZE ZOMER UIT

PARAGRAPH – DESCRIPTION OF GRAPH-RAG PROCEDURE TO IDENTIFY BARRIERS

PARAGRAPH – DESCRIPTION OF GRAPH-RAG PROCEDURE TO IDENTIFY COPING OPTIONS

PARAGRAPH – DESCRIPTION OF CHAIN OF THOUGHT PROCEDURE TO TRANSLATE ALL CURRENT FINDINGS INTO HAPA-BASED TAILORED DIGITAL INTERVENTION

WORDT VERDER UITGEBREID

2.2 Evaluation of Software's Analytical Component

2.2.1 Evaluation of PHOENIX Ontology

All five sub-ontologies will undergo both qualitative and quantitative evaluations. The qualitative assessment consists of two parts: 1) It starts with a technical implementation component (i.e., *process evaluation*) using the OBO Foundry best-practice checklist (Table 1) to confirm adherence to established ontology development standards; and (2) a competency-question (CQ) review (Bezerra et al., 2013; Wilson et al., 2023), in which concise CQ's were crafted to test whether the sub-ontologies do NOT, PARTIALY or FULLY meet certain requirements (i.e., *content evaluation*). For each ontology, twenty diverse, yet representative, use cases will be devised for each CQ that aim to optimally 'stress-test' the ontology. For the quantitative evaluation, all the latency distributions will be visualized of critical ontology-driven operations—such as taxonomy lookups, or cardinality-based ranking of mapping estimates—to ensure that response times remain low enough to. This is an important evaluation dimension for systems with a requirement for low-latency operations—such as just-in-time adaptive interventions, and conversational agents (Abbasian et al., 2024; Haag et al., 2025).

Table 1

The 12 OBO Foundry Principles – Good Practices for Ontology Development

PRINCIPLE	DESCRIPTION
Openness	*The ontology must be freely and publicly accessible online (availability on request is insufficient).*
Standard Formalism	*It must be encoded in a standard ontology format (e.g. OWL/RDF-XML) to ensure tool compatibility.*
Unique Identifiers	*Every class and property must carry a persistent, globally unique identifier (URI) for unambiguous referencing.*
Version Management	*Releases must be clearly labelled with version identifiers, publication dates, and detailed change logs.*
Textual Definitions	*The majority of classes—especially top-level concepts—must include human-readable definitions.*
Naming Conventions	*All terms must follow consistent, intelligible naming rules and be uniquely named.*
Documentation	*Comprehensive supporting materials (e.g. journal articles, online manuals, developer guides) must be available.*
Locus of Authority	*Contact details (name + valid email) for ontology maintainers must be provided.*
Reuse of Existing Resources	*Content should be imported from or mapped to existing ontologies or controlled vocabularies wherever appropriate.*
Documented Plurality of Users	*Evidence of use by multiple independent projects or organizations must be publicly documented.*
Maintenance Plan	*A clear plan for ongoing updates—as well as a record of actual revisions—must be published.*
Community Engagement	*Mechanisms must exist for community feedback, and maintainers must actively address requests.*

Note. In order to evaluate the development and maintenance of a publicly available ontology, a list of good practices was created based on the OBO foundry principles; partially adapted from Smith et al., 2007.

2.2.1.1 'Criterion' Sub-Ontology — Mental Health

A total of ten competency questions have been created to evaluate the performance of the Criterion ontology that aims to represent mental states. Each CQ will be evaluated by a set of 20 representative, yet sufficiently diverse use case scenarios to stress-test the ontologies ability to properly formalize mental health through a sufficiently comprehensive, interoperable, and non-redundant ontology (Table 2).

Table 2

The Ten Competency Questions to Evaluate the Criterion Ontology

ID	Competency Questions
C-01	**Domain Coverage:** Does the ontology enumerate all clinically *and* non-clinically relevant mental-state classes for a specified age group?
C-02	**Hierarchical Structure:** Can any mental-state class be traced through its complete subclass → superclass chain up to the Criterion root?
C-03	**Redundancy-Merging:** Does the system detect and merge symptom labels whose semantic similarity (cos θ) meets or exceeds a set threshold?
C-04	**MR/DCF Accessibility:** For a chosen criterion, are its Mathematical-Restriction and Data-Collection-Feasibility scores directly accessible?
C-05	**Filtering & Ranking:** Can the ontology return—and rank—every criterion satisfying (a) MR-risk $\leq r$, (b) DCF-feasibility $\geq f$, and (c) prevalence $\geq p$ %?
C-06	**Variability Flag:** Are criteria without sufficient intra-day (or sub-hour) variability for data sampling clearly flagged?
C-07	**Mapping Completeness:** Does every leaf criterion have at least one Predictor → Criterion mapping?
C-08	**Semantic Consistency:** Can the ontology surface criterion pairs whose lexical similarity ≥ θ yet live in separate clusters?
C-09	**Text-to-Class:** Given free text (e.g., "I feel anxious and unmotivated"), does the system instantiate the correct criterion classes— in conjunction with the LLM-based approach?
C-10	**Age-Partitioning:** Is each leaf criterion confined to its own age-specific sub-ontology without cross-age leakage?

2.2.1.2 'Predictor' Sub-Ontology — Biopsychosocial Treatment Targets

A total of ten competency questions have been created to evaluate the performance of the Predictor ontology that aims to represent a substantial amount of biopsychosocial treatment variables. Each CQ will be evaluated by a set of 20 representative, yet sufficiently diverse use case scenarios for stress-testing (Table 3). Too much overlap in CQ's between for the Criterion ontology and Predictor ontology is avoided since the developmental

processes to construct the Criterion ontology that have been validated, are likely to also be valid for the Predictor ontology—such as redundancy-related questions and filtering procedures.

Table 3

The Ten Competency Questions to Evaluate the Predictor Ontology

ID	Competency Questions
P-01	**Domain Coverage:** For any top-level biopsychosocial domain, can the ontology list subordinate predictors with sufficient interdisciplinary coverage?
P-02	**Causal Ranking:** For a specified criterion, does it return all predictors with causal relevance ≥ τ, ordered strongest-to-weakest that align with expert-level opinions?
P-03	**Proximal-Zone Flag:** Are predictors outside the user's Proximal Zone of Responsibility (TT-scheme) clearly flagged?
P-04	**Trade-off Retrieval:** Can the system fetch predictors combining high MR-risk with low DCF-feasibility to support trade-off analysis?
P-05	**Gradient Integrity:** Does the ontology consistently demonstrate that all child nodes are a concreter version of their parent node?
P-06	**Path Resolution:** Can it report leaf node predictors that are sufficiently in-depth such that the identification of treatment targets can occur with enough resolution?
P-07	**Cross-Domain Inheritance:** Does it identify predictors inheriting from more than one root domain (Bio, Psycho, Social)?
P-08	**Age-Relevance Gap:** For a given age group, can the ontology surface predictors with age-relevance ≥ σ that are missing from that sub-ontology?
P-09	**Residual Overlap:** Can it detect remaining clusters of semantically overlapping predictors (cos θ ≥ threshold) not yet resolved?
P-10	**Mapping Gaps:** Are there no gaps in which certain predictors lack both Profile → Predictor and Context → Predictor mappings?

2.2.1.3 'Profile' Sub-Ontology — General User Profile

A total of five competency questions have been created to evaluate the performance of the Profile ontology that aims to represent a general profile of the future user. Each CQ will be evaluated by a set of 20 representative, yet sufficiently diverse use case scenarios for stress-testing (Table 4).

Table 4

The Five Competency Questions to Evaluate the Profile Ontology

ID	Competency Questions

PR-01	**Attribute Set:** Does the ontology list relevant profile attributes needed for tailoring digital intervention—demographics, preferences, history, etc.?
PR-02	**Static vs Dynamic:** Are attributes tagged as static (e.g., date of birth) or dynamic (e.g., current stress level)?
PR-03	**Profile-to-Predictor:** Does every profile attribute map to the predictors it modulates over time?
PR-04	**Barrier Alignment:** For each profile and context attribute, can the ontology list the corresponding barrier classes whose likelihood it increases?
PR-05	**Property Rigor:** Are domain, range, and cardinality axioms fully specified for every profile property?

2.2.1.4 'Context' Sub-Ontology — Internal and External Environment

A total of five competency questions have been created to evaluate the performance of the Context ontology that aims to provide a high-resolution overview of contextual factors. Each CQ will be evaluated by a set of 20 representative, yet sufficiently diverse use case scenarios for stress-testing (Table 5).

Table 5

The Five Competency Questions to Evaluate the Context Ontology

ID	Competency Questions
CT-01	**Domain Coverage:** Does the ontology cover all key context domains—time, location, weather, social setting, device, etc.—relevant for tailoring digital interventions?
CT-02	**Context-to-Predictor:** For a selected predictor, can it retrieve contextual factors that modulate its efficacy such that the LLM-based agent can reason more efficiently?
CT-03	**JIT Flags:** Are context variables that mandatory for just-in-time adaptation explicitly flagged given a(ny) identified treatment target?
CT-04	**Context-Barrier Mapping:** For each barrier, can the ontology identify contexts that raise or lower its probability to be relevant for the intervention?
CT-05	**Property Completeness:** Are domain, range, and cardinality constraints declared for every context property?

2.2.1.5 'HAPA-based' Sub-Ontology — Theory-grounded Factors for Behavioral Change

A total of five competency questions have been created to evaluate the performance of the HAPA ontology that aims to provide a high-resolution overview of potential barriers and coping options. Each CQ will be evaluated by a 20 diverse, yet sufficiently diverse use case scenarios for stress-testing (Table 6). The HAPA-based ontology contains two primary sub-ontologies (i.e., for barriers and coping options). For this reason, there might be some replication in the competency questions to evaluate both of them separately.

Table 6

The Five Competency Questions to Evaluate the HAPA Ontology

ID	Competency Questions
H-01	**Barrier Coverage:** For any selected treatment-target predictor, does the ontology enumerate all linked barrier classes—including any subclass or superclass barriers—without omissions?
H-02	**Coping Coverage:** For a specified barrier, can the system retrieve every associated coping strategy that aligns with user opinions?
H-03	**Planning Disjunction:** Are Action-Planning and Coping-Planning constructs explicitly disjoint (separate classes, non-overlapping axioms), so that automated reasoning never conflates them?
H-04	**Barrier-Gap Flag:** Can the ontology automatically surface all barrier classes that currently lack any linked coping strategy, enabling systematic gap analysis and curation?
H-05	**Relation Confidence:** Is each Barrier - Coping relation annotated with confidence or strength score that downstream modules can use for ranking and filtering?

Finally, for all the CQ-queries (incl. other ontologies), latency distributions will be plotted for three different types of query methods that query the stored information to compare what language, and therefore database management system, would be suited for the real-time inference procedures: a) *SPARQL*; B) *Cypher* (Neo4j) and an alternative approach for relational database systems using the query language *PostgreSQL*.

2.2.2 Evaluation of Hierarchical Updating Algorithm

2.1.2.1 Multivariate Time Series Analysis

To evaluate the multivariate time series analysis, a simulation procedure will be employed. In total, 10.000 time-varying latent profiles will be generated with several randomly selected parameter values. The randomly selected parameters that will vary are: 1) the number of variables (from 4 to 20); 2) the amount of datapoints (from 5 to 200); 3) the relational strength between each possible pair; including auto-correlations (from -1 to 1); and 4) the degree of sinusoidal fluctuation to mimic real-life non-stationarity. This last parameter will be drawn from a uniform range between 0 (i.e., no time-varying change) and 1 (i.e., full sinusoidal modulation of the edge strength over time with an amplitude equal to the randomly generated edge strength).

Using these latent time-varying matrices, multivariate timeseries will be simulated with a fixed noise of 0.3 standard deviations using the R package *mgm*. These multivariate signals will then be analyzed by three distinct analysis procedures with the prediction that time-varying gVAR will outperform both stationary estimation methods (i.e., regular gVAR and regular correlations), and the regular correlation method will have the lowest performance. In order to quantify the matrix divergence between the actual matrix content and the predicted matrix content, the Frobenius Norm (Bottcher &Wenzel, 2008) will be computed for each locally

estimated matrix, followed by taking the average of these divergence scores; which is only possible for time-varying models since this is the only procedure where multiple matrices are estimated for each multivariate signal. To evaluate whether there are significant differences in performance between the three analytical methods, a one-way repeated-measures analysis of variance (ANOVA) will be performed on the average of the Frobenius Norms. Assumptions of sphericity will be assessed using Mauchly's test; if violations are detected, then the degrees of freedom will be adjusted accordingly using the Greenhouse- Geisser correction to ensure valid inference. If significant differences are found for the one-way repeated-measure ANOVA, then three paired t-tests will be conducted between all possible model pairs.

2.1.2.2 Quantification of Momentary Impact

To compare the performance of the quantification method with that of a layperson, a short study will be conducted with a Qualtrics-based survey. In this test, 30 Flemish participants will be recruited using social media platforms for a five-minute survey where they will be instructed to rank five variables from least impactful to most impactful; for a total of 10 trials with 30 seconds per trial. In each trial, they will see a picture of three time-varying networks containing five predictors and three criterions with the following instruction: "*Rank each predictor factor (blue circles) from least to most impactful on the criterion factors (purple circles). Note that recent findings are more important than older findings.*" This ancillary study aims to evaluate laypeople's ability to estimate impactful predictors based on raw graphs—which is an important intermediary research question, since modern applications that use intensive longitudinal studies typically only provide raw graphs/matrices instead of direct summary estimates (Onlinehulp-apps.be, 2025).

For this procedure, 10 latent rankings will be manually crafted and reversed-engineered into a noisy time-varying network representation. The performance of the quantification method from the software solution will be compared with the performance of the users by computing Spearman's Footrule—which is a rank-based distance metric that quantifies the absolute deviation between two rankings (Diaconis & Graham, 1977). This metric will be used as dependent variable inside a linear mixed-effects model with *'estimator'* as fixed effect (i.e., quantification method vs user method) and *network_ID* as random effect—including random intercepts and slopes. It is hypothesized that the quantification method will yield superior performance, potentially supporting an initial claim that current approaches used by modern mobile and web applications to communicate about multivariate analysis outputs often lack sufficient interpretability for their end users.

2.2.3 Evaluation of Modular Agentic Framework

The LLM-based agentic framework consists of five sequential modules, each leveraging the capabilities of LLM's to execute reasoning tasks within their local objectives that would otherwise be infeasible using traditional computational methods. Given the technologically exploratory nature of each component within this agentic framework—operationalization, construction of initial model, treatment target identification,

translation of identified targets into tailored intervention, and construction of updated model—, a systematic evaluation of each separate module is vital for assessing their effectiveness and informing further refinement possibilities. Several Qualtrics surveys will be created to compare and evaluate the performance of the LLM-based architectures where both intended non-expert users and several healthcare experts will be recruited. All evaluation procedures will use either ranking methods or scoring methods with a 9-point Likert scale.

Given that (financial) resources are limited, the following research designs try to balance statistical power and internal validity—since multiple studies will be conducted. While each individual study may be on the lower part of the statistical power spectrum, the aim of this multi-study evaluation framework is to gain a detailed understanding of the overall LLM-based reasoning framework by examining each decision point in isolation. Once all the data has been collected, a holistic analysis will be conducted in which all performance-related measures are integrated into a unified linear mixed-effects model—thereby ensuring sufficient statistical power to evaluate the overall performance of the LLM-based reasoning framework.

2.1.3.1 Operationalization of (Non-)Clinical Mental Health Problem

The objective of this module is to operationalize a (textual) description of any mental state into a small set of criterions that are derived from the Criterion ontology. For this reason, a comparative design was created—with feasibility in mind—to compare the operationalization performance of healthcare professionals and the LLM-based reasoning module with each other. In total, 10 healthcare professionals in the mental healthcare industry will be recruited for an online survey. Five of the healthcare professionals will complete a 10-minute survey where they are instructed to write down five labels that would optimally operationalize a description of a (non-)clinical mental state; a short explanation will be provided in the beginning of the survey about the two primary suitability constraints—mathematical restrictions and data collection feasibility issues. A total of 10 diverse, yet representative, textual descriptions will be created beforehand—resulting in 10 trials per survey.

Afterwards, the remaining five healthcare professionals will receive a complementary 20-minute survey where they evaluate the operationalization performance of the previous healthcare professionals and the LLM-based reasoning module—blinded procedure to avoid rating biases. Using the same textual descriptions from first survey, the LLM-based reasoning module also created 10 sets of five labels that stem from the Criterion ontology—resulting in total number of 20 question blocks (i.e., trials). The healthcare professionals will be instructed to evaluate whether the sets of five labels adequately operationalize the description of a mental state on three evaluation dimensions: 1) overall accurate depiction—taking into account an appropriate resolution (i.e., both in breadth and depth); 2) mathematical suitability; 3) data collection feasibility. A short description of the optimization engine will be provided to avoid confusion about the overall objective. For the analysis, three mixed-effects models—separately for each dimension—will be tested where the *'operationalizor'* is used as fixed effect (i.e. expert vs LLM-based reasoner) and the *text_ID* is used as random effect. Bonferroni-correction for multiple comparisons will be applied to avoid issues with an inflating type I error. Given the

technical nature of this evaluation procedure, no ratings will be obtained by the target non-expert users in this first study. This design tries to balance statistical power and internal validity—which is an important consideration when dealing with limited resources, given that multiple studies will be conducted.

2.1.3.2 Construction of Initial Observational Model

The objective of this LLM-based reasoning module is to select a set of predictors from the Predictor ontology with a high plausibility to have a causal relation with any given mental state. To evaluate its performance, a similar design will be used as in the first section where a total number of 10 healthcare professionals will be recruited for an online survey. Once again, five of the healthcare professionals will complete a 10-minute survey, but now they will be instructed to produce a set of five predictors that would complement a set of five criterions—given three suitability constraints (i.e., mathematical restrictions, data collection feasibility, and now also treatment translation). It sounds tempting to just combine this second study with the first study, but if the initial set of generated criterions—by the professionals—were not optimal, this may confound further interpretation. A new set of 10 diverse, yet representative, textual descriptions will be created beforehand—followed by a manually selected set of five criterion labels. This results in 10 sets of five criterion labels (incl. textual description) that will be presented to the healthcare professionals.

Afterwards, the remaining five healthcare professionals will receive a complementary 20-minute survey where they evaluate the performance of the previous healthcare professionals and the LLM-based reasoning module—also a fully blinded procedure to avoid rating biases. The LLM-based framework, also generated these ten sets of five predictor labels using the same input—resulting in total number of 20 question blocks (i.e., trials) for the five healthcare professionals. They will be instructed to evaluate whether the sets of five predictor labels adequately complement the criterion labels to form an observation model for a single-subject longitudinal data collection period. The following four dimensions will be evaluated: 1) overall accurate depiction of potential (causal) predictors—taking into account an appropriate resolution (i.e., both in breadth and depth); 2) mathematical suitability; 3) data collection feasibility; 4) treatment translation possibilities. For the analysis, four separate mixed-effects models will be tested where the *'constructor'* is used as fixed effect (i.e. expert vs LLM-based reasoner) and the *item_ID* is used as random effect. Bonferroni-correction for multiple comparisons will be applied to avoid issues with an inflating type I error. Once again, the technical nature of this evaluation procedure would justify why intended (non-expert) users are not used for this study.

2.1.3.3 Identification of Personalized Treatment Targets

This reasoning module aims to find a subset of predictors based on 1) the MI coefficients and mean values to indicate the current state of the predictors (structured data), 2) three lists containing relevant personal, contextual and barrier information (also structured data), and 3) other textual information that has been collected about the user (unstructured information). To evaluate how well the LLM-based reasoning module can identify treatment targets based on (un)structured personal data, a comparative evaluation design will be

used similar to section 2.1.2.2—where 30 non-expert participants will be recruited for a 20-minute online survey. A total of 10 (non-)clinical use cases will be devised that are reverse engineered into 1) five visualized MI coefficients and visualized mean values that provide information on the current state of those predictors, 2) three lists in which each list contains three relevant short pieces of information, and 3) other information about the pseudo-profile. The participants will be instructed to use this information to rank a total of ten variables (five from observation model + five other potential treatment targets) from having the lowest negative impact to highest negative impact. Note that this differs from the evaluation design in section '2.1.2.2' due to the fact that now 'negative' impact values must be ranked. This is a crucial distinction because strong statistical associations (i.e., high impact scores) alone do not directly imply that these predictors are therefore the most suitable treatment targets; this would only be the case for 'bad' states—based on abnormal mean values that surpass (non-)clinical thresholds across the observation period of those predictors.

Beforehand, 10 latent rankings will be manually crafted to compare the ranking performance of the LLM-based reasoning module and the participants. The performances will be compared by computing Spearman's Footrule between the actual ranking and the predicted ranking—resulting in 600 datapoints (30x10 machine-based performances where the same procedure is repeated 30 times for each task_ID + 30x10 human-based performances where 30 participants perform the ranking procedure for 10 task_ID's). This Footrule will be used as the dependent variable inside a linear mixed-effects model with *'estimator'* as fixed effect (i.e., LLM-based method vs user-based method) and *task_ID* as random effect—including random intercepts and slopes. It is hypothesized that the LLM-based reasoning will yield superior ranking scores—, once again corroborating an initial claim that current approaches used by modern mobile and web applications to communicate complex analysis outputs often lack sufficient interpretability for their end users.

2.1.3.4 Construction of Updated Observational Model

The objective of this LLM-based reasoning module is to update a previous observation model—drawing on recently obtained (un)structured information. Single-subject studies are known to coincide with high attrition rates, which are often directly related to the inherently intensive nature of personal data collection (Schroé et al., 2022). By utilizing a *breadth-first search algorithm,*—which prioritizes abstract entities (until processed) over concrete entities during a targeted search in a hierarchical solution space—this adaptive idiographic modelling approach is expected to result in a superior performance for modelling complex dynamics, while dealing with the problem of wasting resources (e.g., requesting individuals to collect data that is not relevant).

To test this fourth LLM-based reasoning module, a similar design will be employed to that of section 2.1.3.2, in which 10 healthcare professionals will be recruited for an online survey. Five healthcare professionals will be instructed to update five predictor labels from a set of 10 variables (i.e., 5 criterions and 5 predictors). Additionally, a small message will be displayed (underneath the set of variables) about two of those initial predictors being identified as potential treatment targets based on an intensive single-subject longitudinal

study. In the beginning of the survey, a short explanation will be given about the objective of this optimization engine—including the three constraints on the set of predictors (i.e., mathematical restrictions, data collection feasibility, and treatment translation possibility). In total, 10 diverse, yet representative, sets of variables will be created—together with a small message about two identified treatment targets. The LLM-based reasoning module will be provided with the same information so that their updating performances can be compared.

Afterwards, the remaining five healthcare professionals will receive a complementary 20-minute survey where they evaluate the performance of the other healthcare professionals and the LLM-based reasoning module—once again, a fully blinded procedure to avoid rating biases. They will be instructed to evaluate each updated set of predictors on five dimensions: 1) overall accurate depiction of potential (causal) predictors—taking into account an appropriate resolution (i.e., both in breadth and depth); 2) mathematical suitability; 3) feasibility issues related to data collection; 4) treatment translation possibilities; 5) whether the updated predictors align with how the breadth-first search algorithm operates (i.e., either search for other approaches = broad search, or refine the identified predictors by splitting them into relevant sub-predictors). For the analysis, five separate mixed-effects models will be tested where the *'constructor'* is used as fixed effect (i.e. expert vs LLM-based reasoner) and the *task_ID* is used as random effect. Bonferroni-correction for multiple comparisons will be applied to avoid issues with an inflating type I error. Once again, the predominantly technical nature of this evaluation procedure partially justifies why intended (non-expert) users are not included in this design.

2.1.3.5 Translation of Identified Targets into Tailored Digital Intervention

This final LLM-based reasoning module aims to transform identified treatment targets into a concrete action schema. There are many degrees of freedom as to how this communication can occur—but for sake of keeping it feasible, the communication method in this design will be in the form of a short text-only message that can be displayed on a mobile screen. A total of five healthcare expert and 30 non-expert participants will be recruited for a two-part online survey study. The five healthcare professionals will be instructed to generate a short-form message that translates identified treatment targets into a digital intervention. Additionally, short descriptions of relevant personal, contextual and barrier-related information will also be provided to further tailor the digital intervention. This first survey is expected to take up to 20-30 minutes, and contains a total amount of 20 question blocks (i.e., trials). The LLM-based reasoning model will receive the same information and instructions—which then allows for a systematic comparison by recruiting 30 layperson participants.

The participants—in this second complementary survey—will be instructed to evaluate the message content on five dimensions using a 9-point slider: 1) overall congruence; 2.) depth of tailoring; 3) actionability; 4) professional tone; 5) predicted effectiveness. The healthcare professionals that generate the messages will be informed about the utilized scoring matrix by the users. Whether there are any significant differences will be determined by five separate mixed-effects models where the *'generator* is used as fixed effect (i.e. expert vs

LLM) and the *task_ID* is used as random effect. Bonferroni-correction for multiple comparisons will be applied to avoid issues with an inflating type I error. Based on prior findings (Haag et al., 2025) with a similar research design, the LLM-based messages are predicted outperform human-based messages on all domains.

2.1.3.6 Holistic Approach by Integrating all Performance Estimates in a Unified Model

Given that four of the five abovementioned analyses (i.e., 1,2,4, and 5) contain performance estimates of both the LLM-based reasoning module and healthcare professionals, all trial information can be aggregated into a single normalized trial-wise performance estimate. These aggregated performance scores will be used as the dependent variable inside a unifying linear mixed-effects model where the *'reasoner* is used as fixed effect (i.e. healthcare professional vs LLM-based reasoning module) and the *task_ID* is used as random effect. There is no a-priori hypothesis regarding which party will achieve superior performance. However, if the LLM-based reasoning module performs on par with or better than the healthcare professionals, this would provide strong support for the validity of the optimization engine PHOENIX.

2.1.3.7 Simulation-based Approach using Large Language Models as Evaluators

Depending on the remaining time, a large-scale simulation will be executed to identify use-case domains in which the software solution exhibits limited applicability. A total of 100,000 randomized scenarios will be generated by sampling combinations of criteria from the Criterion ontology and converting each into a concise textual description of a (non-)clinical mental state problem—along with (non-)static profile information and (non-)static contextual data. For each use-case scenario, a four-week multivariate time series will be randomly sampled with distinct and realistic statistical patterns. Given the volume of required evaluations, large language models will be employed as evaluators at every decision point (Li et al., 2024). In a pilot study, the statistical association between LLM-generated scores and human ratings (i.e., from healthcare professionals and lay participants) will be computed. If the LLM-generated evaluations demonstrate at least a moderate correlation with human ratings (i.e., $r > 0.3$), their use will be considered justified for large-scale assessment. Given that the observation models will be updated weekly, a total of 400,000 time-varying gVARs will be produced. As the individual use-case results are independent, they can be processed in parallel; this will be carried out using the high-performance computing infrastructure at Ghent University.

The primary objective of this large-scale simulation is to systematically delineate specific use-case domains in which the current implementation of the PHOENIX ontology, multivariate modeling pipeline, or LLM-based modular agentic framework exhibits suboptimal performance, or limited generalizability. The findings will be used to systematically allocate development resources to areas requiring further iteration and precision tuning.

All methods—including study protocols, sampling procedures, analysis plans, and expected outcomes—that will be employed to evaluate 1) the PHOENIX ontology, 2) the multivariate analysis, and 3) the modular agentic framework are pre-registered on Open Science Framework (Foster & Deardorff, 2017): Link1

3. RESULTS

...

4. DISCUSSION

...

5. CONCLUSION

…

6. AI ACKNOWLEDGEMENT SECTION

AI tools, specifically OpenAI's 'o4-mini' model, was used to support the writing process of this report. Its contributions included: 1) rephrasing sentences for clarity and academic tone; 2) generating alternative formulations; 3) summarizing background literature; 4) generating full APA7 references; 5) assisting code generation. All AI-generated content was obviously critically reviewed. The scientific reasoning—including: 1) the formulation of the research question; 2) the literature review, and 3) the methodological choices—are entirely the result of my own critical analysis and independent academic judgment.

7. REFERENCES

Abbasian, M., Khatibi, E., Azimi, I. *et al.* Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *npj Digit. Med. 7*, 82 (2024). https://doi.org/10.1038/s41746-024-01074-z

Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey. *IEEE Access*, *13*, 18912-18936. https://doi.org/10.1109/access.2025.3532853

Agbailu, A. O., Seno, A., & Clement, O. O. (2021). Kalman filter algorithm versus other methods of estimating missing values: Time series evidence. *African Journal of Mathematics and Statistics Studies*, *4*(2), 1-9. https://doi.org/10.52589/ajmss-vfvnmqlx

Akkiraju, R., Xu, A., Bora, D., Yu, T., & An, L. (2024). FACTS About Building Retrieval Augmented Generation-based Chatbots. *arXiv*. https://arxiv.org/abs/2407.07858

Al Khatib, H. S., Neupane, S., Kumar Manchukonda, H., Golilarz, N. A., Mittal, S., Amirlatifi, A., & Rahimi, S. (2024). Patient-centric knowledge graphs: A survey of current methods, challenges, and applications. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1388479

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). *Flamingo: a Visual Language Model for Few-Shot Learning*. arXiv. https://arxiv.org/abs/2204.14198

Alkamli, S., Johnson, L., & Chen, R. (2024). *Ethical and Legal Considerations of Large Language Models: A Systematic Review of the Literature*. ResearchGate. https://www.researchgate.net/publication/388474763

Allsopp, K., Read, J., Corcoran, R., & Kinderman, P. (2019). Heterogeneity in psychiatric diagnostic classification. *Psychiatry Research*, *279*, 15-22. https://doi.org/10.1016/j.psychres.2019.07.005

Altman, A. D., Shapiro, L. A., & Fisher, A. J. (2020). Why does therapy work? An idiographic approach to explore mechanisms of change over the course of psychotherapy using digital assessments. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.00782

Álvarez, A. S., Pagani, M., & Meucci, P. (2012). The clinical application of the biopsychosocial model in mental health: A research critique. *American Journal of Physical Medicine & Rehabilitation*, *91*(13 Suppl 1), S173–S180. https://doi.org/10.1097/PHM.0b013e31823d54be

American Psychiatric Association. (2023). *Understanding mental disorders: Your guide to DSM-5-TR®.* https://doi.org/10.1176/appi.books.9781615375370

Amoretti, M. C., Frixione, M., Lieto, A., & Adamo, G. (2019). Ontologies, mental disorders and prototypes. *Philosophical Studies Series*, 189-204. https://doi.org/10.1007/978-3-030-01800-9_10

Arif, K. H. I., Dip, S. A., Hussain, K., Zhang, L., & Thomas, C. (2025). *PAINT: Paying Attention to INformed Tokens to Mitigate Hallucination in Large Vision-Language Models*. arXiv. https://arxiv.org/abs/2501.12206

Beck, E. D., & Jackson, J. J. (2019). Idiographic traits: A return to Allportian approaches to personality. https://doi.org/10.31234/osf.io/r8xfh

Belani, H., Šolić, P., Zdravevski, E., & Trajkovik, V. (2025). Internet of things Ontologies for well-being, aging and health: A scoping literature review. *Electronics*, *14*(2), 394. https://doi.org/10.3390/electronics14020394

Bezerra, C., Freitas, F., & Santana da Silva, F. (2013). Evaluating ontologies with competency questions. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology* (Vol. 3, pp. 284–287). IEEE. https://doi.org/10.1109/WI-IAT.2013.199

BioPortal. (2024). *BioPortal ontology repository*. National Center for Biomedical Ontology. https://bioportal.bioontology.org/

Bladon, S., Eisner, E., Bucci, S., Oluwatayo, A., Martin, G. P., Sperrin, M., Ainsworth, J., & Faulkner, S. (2025). A systematic review of passive data for remote monitoring in psychosis and schizophrenia. *npj Digital Medicine, 8*, Article 62. https://doi.org/10.1038/s41746-025-01451-2

Böttcher, A., & Wenzel, D. (2008). The Frobenius norm and the commutator. *Linear Algebra and its Applications*, *429*(8-9), 1864-1885. https://doi.org/10.1016/j.laa.2008.05.020

Braun, M., Carlier, S., De Paepe, A., De Backere, F., De Turck, F., & Crombez, G. (2024). Development and evaluation of the contextualised and personalised physical activity and exercise recommendations (COPPER) ontology. https://doi.org/10.31234/osf.io/3pbka

Braun, M., Crombez, G., De Backere, F., Tack, E., & De Paepe, A. (2024). An analysis of physical activity coping plans: Mapping barriers and coping strategies based on user ratings. https://doi.org/10.31234/osf.io/nvg8t

Bringmann, L., Ferrer, E., Hamaker, E., Borsboom, D., & Tuerlinckx, F. (2015). Modeling Nonstationary emotion dynamics in dyads using a Semiparametric time-varying vector autoregressive model. *Multivariate Behavioral Research*, *50*(6), 730-731. https://doi.org/10.1080/00273171.2015.1120182

Broniatowski, M., & Keziou, A. (2006). Minimization of φ-divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica*, *43*(4), 403-442. https://doi.org/10.1556/sscmath.43.2006.4.2

Casella, A., & Wang, W. (2025). *Performant LLM Agentic Framework for Conversational AI*. arXiv preprint arXiv:2503.06410. https://doi.org/10.48550/arXiv.2503.06410

Castro, D., Gysi, D., Ferreira, F., Ferreira-Santos, F., & Ferreira, T. B. (2024). Centrality measures in psychological networks: A simulation study on identifying effective treatment targets. *PLOS ONE*, *19*(2), e0297058. https://doi.org/10.1371/journal.pone.0297058

Cesa-Bianchi, N., Gentile, C., Lugosi, G., & Neu, G. (2017). Boltzmann exploration done right*.), Advances in Neural Information Processing Systems, 30* (pp. 6284–6293). Curran Associates, Inc. https://doi.org/10.48550/arXiv.1705.10257

Christensen, A. P., Choi, J., & Garrido, L. E. (2024). Evaluating network Replicability across local, Mesoscale, and global structures. https://doi.org/10.31234/osf.io/9d4nm

Cuthbert, B. N. (2020). The role of RDoC in future classification of mentaldisorders. *Dialogues in Clinical Neuroscience*, *22*(1), 81-85. https://doi.org/10.31887/dcns.2020.22.1/bcuthbert

Dang, H. T. (2006). *Overview of DUC 2006*. In *Proceedings of the Document Understanding Conference*. https://duc.nist.gov/pubs/2006papers/duc2006.pdf

Danner, D., et al. (2023). *Advancing depression detection on social media platforms through fine-tuned large language models*. arXiv. https://arxiv.org/abs/2409.14794

De Angel, V., Lewis, S., White, K., Oetzmann, C., Leightley, D., Oprea, E., Lavelle, G., Matcham, F., Pace, A., Mohr, D. C., Dobson, R., & Hotopf, M. (2022). Digital health tools for the passive monitoring of depression: A systematic review of methods. *npj Digital Medicine*, 5, 3. https://doi.org/10.1038/s41746-021-00548-8

De mythes bevraagd. Resultaten van de public mental health monitor 2023. (2024, June 2). Zorgnet-Icuro. https://www.zorgneticuro.be/publicaties/de-mythes-bevraagd-resultaten-van-de-public-mental-health-monitor-2023

Diaconis, P., & Graham, R. L. (1977). Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *39*(2), 262-268. https://doi.org/10.1111/j.2517-6161.1977.tb01624.x

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., ... & Cui, C. (2022). *GLaM: Efficient Scaling of Language Models with Mixture-of-Experts*. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 5547–5569). PMLR. https://proceedings.mlr.press/v162/du22c.html

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., & Larson, J. (2024). *From Local to Global: A Graph RAG Approach to Query-Focused Summarization*. arXiv. https://arxiv.org/abs/2404.16130

Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. *Science*, 196(4286), 129–136. https://doi.org/10.1126/science.847460

Elekes E., M. Schaeler and K. Boehm, On the Various Semantics of Similarity in Word Embedding Models. *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Toronto, ON, Canada, 2017, pp. 1-10, https://doi.org/10.1109/JCDL.2017.7991568

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, *53*(4), 453-480. https://doi.org/10.1080/00273171.2018.1454823

Erikson, E. H. (1950). Childhood and society (2nd Ed.). New York: Norton. https://archive.org/details/dli.ernet.19961

Explainability for large language models: A survey. arXiv. https://arxiv.org/abs/2309.01029

Fanali, A., Giorgi, F., & Tramonti, F. (2024). Thick description and systems thinking: Reiterating the importance of a biopsychosocial approach to mental health. *Journal of Evaluation in Clinical Practice*. Advance online publication. https://doi.org/10.1111/jep.13800

Farahani, F. V., Karwowski, W., & Lighthall, N. R. (2019). Application of graph theory for identifying connectivity patterns in human brain networks: A systematic review. *Frontiers in Neuroscience*, *13*. https://doi.org/10.3389/fnins.2019.00585

Fedus, W., Zoph, B., & Shazeer, N. (2021). *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*. arXiv. https://arxiv.org/abs/2101.03961

Ferraris, A. F., Audrito, D., Caro, L. D., & Poncibò, C. (2025). The architecture of language: Understanding the mechanics behind LLMs. *Cambridge Forum on AI: Law and Governance*, *1*. https://doi.org/10.1017/cfl.2024.16

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*(27). https://doi.org/10.1073/pnas.1711978115

Forbes, M. K., Neo, B., Nezami, O. M., Fried, E. I., Faure, K., Michelsen, B., Twose, M., & Dras, M. (2023). Elemental psychopathology: Distilling constituent symptoms and patterns of repetition in the diagnostic criteria of the DSM-5. *Psychological Medicine*, *54*(5), 886-894. https://doi.org/10.1017/s0033291723002544

Foster, E. D., & Deardorff, A. (2017). Open science framework (OSF). *Journal of the Medical Library Association*, *105*(2). https://doi.org/10.5195/jmla.2017.88

Frumkin, M., Piccirillo, M., Beck, E. D., Grossman, J., & Rodebaugh, T. (2019). Feasibility and utility of idiographic models in the clinic: A pilot study. https://doi.org/10.31234/osf.io/m34aw

Gaber, F., Shaik, M., Franke, V., & Akalin, A. (2025). Evaluating large language model workflows in clinical decision support: Referral, triage, and diagnosis. https://doi.org/10.1101/2024.09.27.24314505

Galenos. (2024). *Galenos ontology repository*. University of Manchester. https://www.galenos.org.uk/ontology

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). *Retrieval-Augmented Generation for Large Language Models: A Survey*. arXiv. https://arxiv.org/abs/2312.10997

Gennari, J. H., Musen, M. A., Fergerson, R. W., Grosso, W. E., Crubézy, M., Eriksson, H., Noy, N. F., & Tu, S. W. (2003). *The evolution of Protégé: An environment for knowledge-based systems development*. International Journal of Human-Computer Studies, 58(1), 89–123. https://doi.org/10.1016/S1071-5819(02)00127-1

GNU GENERAL PUBLIC LICENSE. (2007). Free Software Foundation. https://www.gnu.org/licenses/gpl-3.0.nl.html

Governance actieplan eGezondheid. (2025). Rijksinstituut voor ziekte- en invaliditeitsverzekering. https://www.riziv.fgov.be/nl/thema-s/egezondheid/interfederaal-actieplan-egezondheid-2025-2027

Großwendt, A., Röglin, H. Improved Analysis of Complete-Linkage Clustering. Algorithmica 78, 1131–1150 (2017). https://doi.org/10.1007/s00453-017-0284-6

Grymonprez, S. (2018). Psychische problemen kosten ons jaarlijks 20 miljard. *De Standaard*. https ://www.standaard.be/cnt/dmf20181122_03975492

Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). Large language models for mental health applications: Systematic review. *JMIR Mental Health*, *11*, e57400. https://doi.org/10.2196/57400

Haag, D., Kumar, D., Gruber, S., Hofer, D. P., Sareban, M., Treff, G., Niebauer, J., Bull, C. N., Schmidt, A., & Smeddinck, J. D. (2025). The last JITAI? Exploring large language models for issuing just-in-Time adaptive interventions: Fostering physical activity in a prospective cardiac rehabilitation setting. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1-18. https://doi.org/10.1145/3706598.3713307

Han, Z., Gao, C., Liu, J., Zhang, J., & Zhang, S. Q. (2024). *Parameter-efficient fine-tuning for large models: A comprehensive survey*. arXiv. https://arxiv.org/abs/2403.14608

Haslbeck, J. M., & Waldorp, L. J. (2020). MGM: Estimating time-varying mixed graphical models in high-dimensional data. *Journal of Statistical Software*, *93*(8). https://doi.org/10.18637/jss.v093.i08

Hayou, S., Ghosh, N., & Yu, B. (2024). *LoRA+: Efficient Low Rank Adaptation of Large Models*. arXiv. https://arxiv.org/abs/2402.12354

Henry, L. M., Hansen, E., Chimoff, J., Pokstis, K., Kiderman, M., Naim, R., Kossowsky, J., Byrne, M. E., Lopez-Guzman, S., Kircanski, K., Pine, D. S., & Brotman, M. A. (2024). Selecting an ecological momentary assessment platform: Tutorial for researchers. *Journal of Medical Internet Research*, *26*, e51125. https://doi.org/10.2196/51125

Hoekstra, R. H., De Ron, J., Epskamp, S., Robinaugh, D., & Borsboom, D. (2024). Mapping the dynamics of idiographic network models to the network theory of psychopathology using stability landscapes. https://doi.org/10.31234/osf.io/9sguw

Holland AM, Bon-Frauches AC, Keszthelyi D, Melotte V, Boesmans W. The enteric nervous system in gastrointestinal disease etiology. Cell Mol Life Sci. 2021 May;78(10):4713-4733. https://doi.org/10.1007/s00018-021-03812-y

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv. https://arxiv.org/abs/2106.09685

Hua, Y., Liu, F., Yang, K., Li, Z., Sheu, Y., Zhou, P., Moran, L. V., Ananiadou, S., Clifton, D. A., & Beam, A. (2024). Large language models in mental health care: A systematic scoping review (Preprint). https://doi.org/10.2196/preprints.64088

Hua, Y., Na, H., Li, Z., Liu, F., Fang, X., Clifton, D., & Torous, J. (2025). A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine*, *8*(1). https://doi.org/10.1038/s41746-025-01611-4

Huang, Q., & Zhao, T. (2024). *Leveraging large language models for entity matching*. arXiv. https://arxiv.org/abs/2405.20624

Huang, Y., Wang, N., Zhang, Z., Liu, H., Fei, X., Wei, L., & Chen, H. (2021). Patient representation from structured electronic medical records based on embedding technique: Development and validation study. *JMIR Medical Informatics, 9*(7), e19905. https://doi.org/10.2196/19905

Hulsmans, D., Oude Maatman, F., Otten, R., Poelen, E., & Lichtwarck-Aschoff, A. (2024). Idiographic personality networks: Stability, variability and when they become problematic. https://doi.org/10.31234/osf.io/xf65q

ICD-11 2024 release. (2024, February 8). World Health Organization (WHO). https://www.who.int/news/item/08-02-2024-icd-11-2024-release

Jackson, R., Matentzoglu, N., Overton, J. A., Vita, R., Balhoff, J. P., Buttigieg, P. L., Carbon, S., Courtot, M., Diehl, A. D., Dooley, D. M., Duncan, W. D., Harris, N. L., Haendel, M. A., Lewis, S. E., Natale, D. A., Osumi-Sutherland, D., Ruttenberg, A., Schriml, L. M., Smith, B., … Peters, B. (2021). OBO foundry in 2021: Operationalizing open data principles to evaluate ontologies. *Database*, *2021*. https://doi.org/10.1093/database/baab069

Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G. (2023). *Active Retrieval Augmented Generation*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*(pp. 7969–7992). Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-main.495/

Kaiser, T., & Laireiter, A. (2018). Process-symptom-bridges in psychotherapy: An idiographic network approach. https://doi.org/10.31234/osf.io/xukgm

Kamath, U., Graham, K. L., & Emara, W. (2022). Bidirectional encoder representations from transformers (BERT). *Transformers for Machine Learning*, 43-70. https://doi.org/10.1201/9781003170082-3

Karcioglu O, Topacoglu H, Dikme O, Dikme O. A systematic review of the pain scales in adults: Which to use? Am J Emerg Med. 2018 Apr;36(4):707-714. https://doi.org/10.1016/j.ajem.2018.01.008

Kim, J. H. (2014). Testing for parameter restrictions in a stationary VAR model: A bootstrap alternative. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2395806

Kip, H., Beerlage-de Jong, N., van Gemert-Pijnen, L. J. E. W. C., & Kelders, S. M. (2025). The CeHRes Roadmap 2.0: Update of a holistic framework for development, implementation, and evaluation of eHealth technologies. *Journal of Medical Internet Research, 27*, e59601. https://doi.org/10.2196/59601

Klooster, I. T., Kip, H., van Gemert-Pijnen, L., Crutzen, R., & Kelders, S. M. (2024). A systematic review on eHealth technology personalization approaches. *iScience, 27*, 110771. https://doi.org/10.1016/j.isci.2024.110771

Krishna, K., Portsmouth, L., Harris, C., & Ciccarelli, M. (2025). What's the 'Secret sauce'?: A systematic review of the characteristics of effective digital health behaviour change interventions for children and adolescents. *Health Promotion Journal of Australia*, *36*(3). https://doi.org/10.1002/hpja.70051

Kuper, N., Andresen, P. K., Beck, E. D., Costantini, G., Hamaker, E., Wright, A. G., & Zimmermann, J. (2024). From persons to general principles: Methodological decisions for idiographic and nomothetic research. https://doi.org/10.31234/osf.io/mx47u

Lai, V., Ngo, N., Pouran Ben Veyseh, A., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. (2023). ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. *Findings of the Association for Computational Linguistics: EMNLP 2023*. https://doi.org/10.18653/v1/2023.findings-emnlp.878

Larsen, K. R., Michie, S., Hekler, E. B., Gibson, B., Spruijt-Metz, D., Ahern, D., Cole-Lewis, H., Ellis, R. J., Hesse, B., Moser, R. P., & Yi, J. (2016). Behavior change interventions: The potential of ontologies for advancing science and practice. *Journal of Behavioral Medicine*, *40*(1), 6-22. https://doi.org/10.1007/s10865-016-9768-0

Larsen, R. R., & Hastings, J. (2018). From affective science to psychiatric disorder: Ontology as a semantic bridge. *Frontiers in Psychiatry*, *9*. https://doi.org/10.3389/fpsyt.2018.00487

Lawrence, H. R., Schneider, R. A., Rubin, S. B., Matarić, M. J., McDuff, D. J., & Jones Bell, M. (2024). The opportunities and risks of large language models in mental health. *JMIR Mental Health*, *11*, e59479-e59479. https://doi.org/10.2196/59479

Levinson, C. A., Christian, C., & Becker, C. B. (2024). How idiographic methodologies can move the clinical-science Field forward to integrate personalized treatment into everyday clinical care and improve treatment outcomes. *Clinical Psychological Science*, *13*(1), 69-82. https://doi.org/10.1177/21677026231217316

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. arXiv. https://arxiv.org/abs/2005.11401

Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., & Liu, Y. (2024). *LLMs-as-Judges: A comprehensive survey on LLM-based evaluation methods*. arXiv. https://doi.org/10.48550/arXiv.2412.05579

Liu, B., Li, X., Zhang, J., Wang, J., He, T., Hong, S., Liu, H., Zhang, S., (2025). Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. arXiv. https://arxiv.org/abs/2504.01990

Liu, J., Tang, P., Wang, W., Ren, Y., Hou, X., Heng, P.-A., Guo, M., & Li, C. (2024). A survey on inference optimization techniques for mixture of experts models. arXiv. https://arxiv.org/abs/2412.14219

Lunansky, G., Naberman, J., Van Borkulo, C. D., Chen, C., Li, W., & Borsboom, D. (2021). Intervening on psychopathology networks: Evaluating intervention targets through simulations. https://doi.org/10.31234/osf.io/sqhje

Luschi, A., Petraccone, C., Fico, G., Pecchia, L., & Iadanza, E. (2023). Semantic Ontologies for complex healthcare structures: A scoping review. *IEEE Access*, *11*, 19228-19246. https://doi.org/10.1109/access.2023.3248969

Matentzoglu, N., Ronzano, F., Nentwig, M., Pesquita, C., Slater, L., & Jansen, L. (2023). *MapperGPT: Large language models for linking and mapping entities*. arXiv. https://arxiv.org/abs/2310.03666

McCulloch, J. A., St. Pierre, S. R., Linka, K., & Kuhl, E. (2024). On sparse regression, regularization, and automated model discovery. *International Journal for Numerical Methods in Engineering*, *125*(14). https://doi.org/10.1002/nme.7481

Mental health. (2019, December 19). World Health Organization (WHO). https://www.who.int/health-topics/mental-health#tab=tab_1

Mink, F., Lutz, W., & Hehlmann, M. I. (2025). Ecological momentary assessment in psychotherapy research: A systematic review. *Clinical Psychology Review*, *117*, 102565. https://doi.org/10.1016/j.cpr.2025.102565

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201-218. https://doi.org/10.1207/s15366359mea0204_1

Morris, S. E., Sanislow, C. A., Pacheco, J., Vaidyanathan, U., Gordon, J. A., & Cuthbert, B. N. (2022). Revisiting the seven pillars of RDoC. *BMC Medicine*, *20*(1). https://doi.org/10.1186/s12916-022-02414-0

Network analysis of multivariate data in psychological science. (2021). *Nature Reviews Methods Primers*, *1*(1). https://doi.org/10.1038/s43586-021-00060-z

Nicoară, R.-D., Tegzeșiu, A. M., & Popescu, C. A. (2024). *Systematic review of CBT techniques and their alignment with the Transtheoretical Model stages of change*. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 15(4), 253–264. https://doi.org/10.70594/brain/15.4/17

Olbert, C. M., Gala, G. J., & Tupler, L. A. (2014). Quantifying heterogeneity attributable to polythetic diagnostic criteria: Theoretical framework and empirical application. *Journal of Abnormal Psychology*, *123*(2), 452-462. https://doi.org/10.1037/a0036068

Onlinehulp-apps. (2025). *onlinehulp-apps*. onlinehulp-apps. https://www.onlinehulp-apps.be

Ontobee. (2024). *Ontobee ontology repository*. He Group, University of Michigan Medical School. https://www.ontobee.org/

OpenAI. (2023). *Function Calling in OpenAI Models: A Practical Guide*. Towards AI. https://towardsai.net/p/l/the-llm-series-2-function-calling-in-openai-models-a-practical-guide

Orenstein, G. A., & Lewis, L. (2022, November 7). Erikson's stages of psychosocial development. In *StatPearls*. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK556096/

Owen, J., Smith, A., & Lee, M. (2024). *Understanding the Epistemic Opacity of Transformer Architectures*. Journal of Artificial Intelligence Research, 75(1), 123–145. https://doi.org/10.1613/jair.1.13000

Park, J. J., Chow, S., Epskamp, S., & Molenaar, P. (2022). Subgrouping with chain graphical VAR models. https://doi.org/10.31234/osf.io/u3ve8

Peeters, M., Zhang, Y., & Sarkar, P. (2024). *Entity matching using large language models*. arXiv. https://arxiv.org/abs/2310.11244

Petukhova, A., Matos-Carvalho, J. P., & Fachada, N. (2024). *Text clustering with LLM embeddings*. arXiv preprint arXiv:2403.15112. https://arxiv.org/abs/2403.15112

Piccirillo, M. L., Beck, E. D., & Rodebaugh, T. L. (2019). A clinician's primer for idiographic research: Considerations and recommendations. *Behavior Therapy*, *50*(5), 938-951. https://doi.org/10.1016/j.beth.2019.02.002

Pinto, H. S., & Martins, J. P. (2004). A methodology for ontology integration. *Proceedings of the international conference on Knowledge capture - K-CAP 2001*. https://doi.org/10.1145/500742.500759

Prochaska, J. O., & DiClemente, C. C. (1983). *Stages and processes of self-change of smoking: Toward an integrative model of change*. Journal of Consulting and Clinical Psychology, 51(3), 390–395. https://doi.org/10.1037/0022-006X.51.3.390

Pullman, J., Molloy, L., Beckett, P. J., & Campbell, S. (2025). Social work, psychiatry, the biopsychosocial model, and mental health reform. *Social Work in Mental Health*, 23(1), 1–17. https://doi.org/10.1080/15332985.2024.2393872

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Raiaan, M. A., Mukta, M. S., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M., Ahmad, J., Ali, M. E., & Azam, S. (2023). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. https://doi.org/10.36227/techrxiv.24171183.v1

Rathle. (2024, April 25). *The GraphRAG manifesto: Adding knowledge to GenAI*. Graph Database & Analytics. https://neo4j.com/blog/genai/graphrag-manifesto/

Rayhan, M. A., & Ashrafuzzaman, M. (2025). *LLM Enhancer: Merged approach using vector embedding for hallucination mitigation in large language models*. arXiv preprint arXiv:2504.21132. https://arxiv.org/abs/2504.21132

Renze, M. (2024). *The Effect of Sampling Temperature on Problem Solving in Large Language Models. Findings of the Association for Computational Linguistics: EMNLP 2024.* https://www.researchgate.net/publication/386198408

Reyes Fernández, B., Fleig, L., Godinho, C. A., Montenegro, E., Knoll, N., & Schwarzer, R. (2015). *Action control bridges the planning–behavior gap: A longitudinal study on physical exercise in young adults*. Psychology & Health, 30(8), 945–958. https://doi.org/10.1080/08870446.2015.1006222

Rijcken, E., Zervanou, K., Mosteiro, P., Scheepers, F., Spruit, M., & Kaymak, U. (2025). Machine learning vs. rule-based methods for document classification of electronic health records within mental health care: A systematic literature review. *Natural Language Processing*, 10, 100129. https://doi.org/10.1016/j.nlp.2025.100129

Rivera-Romero, O., Gabarron, E., Ropero, J., & Denecke, K. (2023). Designing personalised mHealth solutions: An overview. *Journal of Biomedical Informatics, 146*, 104500. https://doi.org/10.1016/j.jbi.2023.104500

Robinson, C.L., Phung, A., Dominguez, M. et al. Pain Scales: What Are They and What Do They Mean. Curr Pain Headache Rep 28, 11–25 (2024). https://doi.org/10.1007/s11916-023-01195-2

Rocchi, G., Vocaj, E., Moawad, S., Antonucci, A., Grigioni, C., Giuffrida, V., & Bordini, J. (2024). Optimizing personalized psychological well-being interventions through digital phenotyping: Results from a randomized non-clinical trial. *Frontiers in Psychology, 15*, Article 1479269. https://doi.org/10.3389/fpsyg.2024.1479269

Rogan, J., Bucci, S., & Firth, J. (2024). Health care professionals' views on the use of passive sensing, AI, and machine learning in mental health care: Systematic review with meta-synthesis. *JMIR Mental Health, 11*, e49577. https://doi.org/10.2196/49577

Rouvere, J., Blanchard, B. E., Johnson, M., Griffith Fillipo, I., Mosser, B., Romanelli, M., Nguyen, T., Rushton, K., Marion, J., Althoff, T., Areán, P. A., & Pullmann, M. D. (2024). *Application of an adapted Health Action Process Approach model to predict engagement with a digital mental health website: Cross-sectional study*. JMIR Human Factors, 11, e57082. https://doi.org/10.2196/57082

Ryan, O., Haslbeck, J. M., & Waldorp, L. J. (2025). Non-stationarity in time-series analysis: Modeling stochastic and deterministic trends. *Multivariate Behavioral Research*, 1-33. https://doi.org/10.1080/00273171.2024.2436413

Sajnovic, A., Bjelica, M., & Milinkovic, D. (2024). Internet of Things and Big Data Analytics in Preventive Healthcare: A Synthetic Review. *Electronics*, *13*(18), 3642. https://doi.org/10.3390/electronics13183642

Saleem, M., Kühne, L., De Santis, K. K., Brand, T., & Busse, H. (2021). Effective engagement strategies in digital interventions for mental health promotion: A scoping review. *PsychArchives*. https://doi.org/10.23668/psycharchives.4835

Sankar, E., & Dimitri, V. (2025). Mixture of experts models in deep learning and their techniques applications and challenges. https://doi.org/10.36227/techrxiv.174002480.07096488/v1

Schenk, P. M., Hastings, J., Santilli, M., Potts, J., Kennett, J., Friedrich, C., & Michie, S. (2024). Towards an ontology of mental health: Protocol for developing an ontology to structure and integrate evidence regarding anxiety, depression and psychosis. *Wellcome Open Research*, *9*, 40. https://doi.org/10.12688/wellcomeopenres.20701.2

Schroé, H., Van Dyck, D., De Paepe, A., Poppe, L., Verloigne, M., & De Bourdeaudhuij, I. (2022). Investigating when, which, and why users stop using a digital health intervention to promote an active lifestyle: A focus on HAPA-based psychological determinants. *JMIR mHealth and uHealth, 10*(1), e30583. https://doi.org/10.2196/30583

Schwarzbach, N. R., Hoekstra, R., Poppe, A., Bouman, T. K., & Pijnenborg, G. H. (2025). When theory and therapy part ways—A scoping review of the science-to-practice gap. *Psychotherapy Research*, 1-21. https://doi.org/10.1080/10503307.2025.2488019

Schwarzer, R. (2008). *Modeling health behavior change: How to predict and modify the adoption and maintenance of health behaviors*. Applied Psychology, 57(1), 1–29. https://doi.org/10.1111/j.1464-0597.2007.00325.x

Schwarzer, R., & Luszczynska, A. (2008). How to overcome health-compromising behaviors. *European Psychologist*, *13*(2), 141-151. https://doi.org/10.1027/1016-9040.13.2.141

Serafim, P. H., De Sousa, M. H., & Czepielewski, L. S. (2025). Network analysis in psychopathology: Theoretical perspectives and practical challenges. *Trends in Psychology*. https://doi.org/10.1007/s43076-025-00438-y

Siepe, B. S., Sander, C., Schultze, M., Kliem, A., Ludwig, S., Hegerl, U., & Reich, H. (2024). Time-varying network models for the temporal dynamics of depressive symptomatology in patients with depressive disorders: Secondary analysis of longitudinal observational data. *JMIR Mental Health*, *11*, e50136. https://doi.org/10.2196/50136

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S., Scheuermann, R. H., Shah, N., Whetzel, P. L., & Lewis, S. (2007). The OBO foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, *25*(11), 1251-1255. https://doi.org/10.1038/nbt1346

Steck, H., Ekanadham, C., & Kallus, N. (2024). *Is cosine-similarity of embeddings really about similarity?* arXiv. https://arxiv.org/abs/2403.05440

Steiger, E., & Kroll, L. E. (2023). Patient embeddings from diagnosis codes for health care prediction tasks: Pat2Vec machine learning framework. *JMIR AI, 2*(1), e40755. https://doi.org/10.2196/40755

Sukhbaatar, S., Golovneva, O., Sharma, V., Xu, H., Lin, X. V., Rozière, B., ... & Li, X. (2024). *Branch-Train-MiX: Mixing Expert LLMs into a Mixture-of-Experts LLM*. arXiv. https://arxiv.org/abs/2403.07816

Supplemental material for a tutorial on regularized partial correlation networks. (2018). *Psychological Methods*. https://doi.org/10.1037/met0000167.supp

Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., Mathur, P., Cacciamani, G. E., Sun, C., Peng, Y., & Wang, Y. (2024). *A framework for human evaluation of large language models in healthcare derived from literature review*. NPJ Digital Medicine, 7(1), 258. https://doi.org/10.1038/s41746-024-01258-7

Tang, Y., Wang, L., & Zhu, D. (2024). *Extending embedding models for long context retrieval*. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 4375–4391). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.47.pdf

Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). *A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models*. arXiv. https://arxiv.org/abs/2401.01313

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). *From Louvain to Leiden: Guaranteeing Well-Connected Communities*. *Scientific Reports*, *9*(1), 5233. https://doi.org/10.1038/s41598-019-41695-z

Van Veen, M., et al. (2024). *Large language models are poor clinical decision-makers: A comprehensive benchmark*. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 759. https://aclanthology.org/2024.emnlp-main.759.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*(pp. 6000–6010). Curran Associates, Inc. https://arxiv.org/abs/1706.03762

Vavekanand, R., & Kumar, S. (2024). LLMEra: Impact of large language models. https://doi.org/10.2139/ssrn.4857084

Vidaurre, D., Bielza, C., & Larrañaga, P. (2013). A survey of *L1* Regression. *International Statistical Review*, *81*(3), 361-387. https://doi.org/10.1111/insr.12023

Wang, X., Sen, P., Li, R., & Yilmaz, E. (2024). Adaptive retrieval-augmented generation for conversational systems. In *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 655–661). Association for Computational Linguistics. https://aclanthology.org/2025.findings-naacl.30/

Wilson, N., Keet, C. M., & Casmod Khan, Z. (2023). Discerning and characterising types of competency questions for ontologies. *arXiv preprint arXiv:2412.13688*. https://arxiv.org/abs/2412.13688

Woll, S., Birkenmaier, D., Biri, G., Nissen, R., Lutz, L., Schroth, M., Ebner-Priemer, U. W., & Giurgiu, M. (2025). Applying AI in the context of the association between device-based assessment of physical activity and mental health: Systematic review. https://doi.org/10.2196/59660

Wright, A. G., & Woods, W. C. (2020). Personalized models of psychopathology. https://doi.org/10.31234/osf.io/6hqzj

Xu, L., Xie, H., Qin, S.-Z. J., Tao, X., & Wang, F. L. (2023). *Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment*. arXiv. https://arxiv.org/abs/2312.12148

Xuan, H., Yang, B., & Li, X. (2025). Exploring the Impact of Temperature Scaling in Softmax for Classification and Adversarial Robustness. arXiv preprint arXiv:2502.20604. https://arxiv.org/abs/2502.20604

Yamada, D. B., Bernardi, F. A., Miyoshi, N. S., De Lima, I. B., Vinci, A. L., Yoshiura, V. T., & Alves, D. (2020). Ontology-based inference for supporting clinical decisions in mental health. *Lecture Notes in Computer Science*, 363-375. https://doi.org/10.1007/978-3-030-50423-6_27

Ye, H., Liu, T., Zhang, A., Hua, W., & Jia, W. (2023). *Cognitive Mirage: A Review of Hallucinations in Large Language Models*. arXiv. https://arxiv.org/abs/2309.06794

Yuan, Y., et al. (2025). *Artificial intelligence conversational agents in mental health: Patients' perceptions and acceptability*. Frontiers in Psychiatry, 15, 1505024. https://doi.org/10.3389/fpsyt.2024.1505024

Zhang, Z.-R., Tan, C., Xu, H., Wang, C., Huang, J., & Huang, S. (2023). *Towards Adaptive Prefix Tuning for Parameter-Efficient Language Model Fine-tuning*. arXiv. https://arxiv.org/abs/2305.15212

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2023).

Zhao, Q., Nooner, K. B., Tapert, S. F., Adeli, E., Pohl, K. M., Kuceyeski, A., & Sabuncu, M. R. (2025). The transition from homogeneous to heterogeneous machine learning in neuropsychiatric research. *Biological Psychiatry Global Open Science*, *5*(1), 100397. https://doi.org/10.1016/j.bpsgos.2024.100397

Zheng, Z., Liao, L., Deng, Y., Lim, E.-P., Huang, M., & Nie, L. (2024). *Thoughts to target: Enhance planning for target-driven conversation*. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 21108–21124). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.1175

Zhu, Y., Wang, X., Chen, J., Qiao, S., Ou, Y., Yao, Y., Deng, S., Chen, H., & Zhang, N. (2024). LLMs for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web*, *27*(5). https://doi.org/10.1007/s11280-024-01297-w

Zuidersma, M., Riese, H., Snippe, E., Booij, S. H., Wichers, M., & Bos, E. H. (2020). Single-subject research in psychiatry: Facts and fictions. *Frontiers in Psychiatry*, *11*. https://doi.org/10.3389/fpsyt.2020.539777

8. APPENDIX