Psychology-Grounded Individualized Artificial Mental Model

Prajvi Saxena[1], Arvind Nagarajan[1], Sabine Janzen[1], and Wolfgang Maass[1,2]

[1] German Research Center for Artificial Intelligence (DFKI)
[2] Saarland University
prajvi.saxena@dfki.de

Abstract. Personalizing rehabilitation requires predicting how individual patients will experience pain yet two patients undergoing identical surgery can report starkly different outcomes. This heterogeneity is not noise: it is systematically driven by personality, beliefs, emotional state, and life circumstances. Psychology has spent decades formalizing precisely these individual differences through validated frameworks such as the Biopsychosocial model and the International Classification of Functioning. Yet existing AI approaches ignore this theoretical structure and treating patient variables as flat numerical features. We propose Psychology-Grounded Artificial Mental Models: a framework that encodes established psychological theory as structured reasoning scaffolds within large language models, enabling them to construct an artificial mental model of each patient before making predictions. To evaluate the framework, we conducted a DiscoverYourself study with 336 participants, and further validated the framework on publicly available PhysioPain dataset (N=82). Across four LLM families, theory-grounded models achieve up to +56.9 percentage points exact-accuracy improvement over the generic baseline. These results indicate that encoding psychological theory into LLM reasoning substantially improves pain prediction in rehabilitation, with the largest gains observed among older, highly heterogeneous patients where data-driven models are least reliable. Our code repository is available at [3].

Keywords: Artificial Mental Models · Psychology-Grounded AI · Individualized Prediction · Large Language Models · Rehabilitation · Pain Assessment

1 Introduction

Why does the same exercise feel effortless to one patient and agonizing to another even when both underwent the same surgery, share similar demographics, and follow the same rehabilitation protocol? This question lies at the heart of personalized rehabilitation, and its answer is deeply rooted in individual psychology. A patient's experience of pain and exertion is not a straightforward

[3] GitHub:https://anonymous.4open.science/r/anon-repo-0661/README.md

readout of tissue status; it is filtered through layers of personality, emotional regulation, prior beliefs about recovery, lifestyle habits, and social context [6,20]. As populations age and rehabilitation demand grows, the ability to predict these *individual* differences not just population averages becomes increasingly critical for effective, patient-centered care [22]. Psychology has investigated this problem for decades producing well-validated theoretical frameworks that explain individual differences in pain and health outcomes. The Biopsychosocial model (BPS) [5] positions every patient at the intersection of biological, psychological, and social forces. The International Classification of Functioning, Disability and Health (ICF) [24] provides a structured taxonomy of body functions, activities, participation, and contextual factors. The Fear-Avoidance model [15] explains how pain-related fear and catastrophizing drive disability beyond what injury alone predicts. Cognitive-Behavioral frameworks [3] identify how maladaptive thought patterns sustain pain perception long after tissue healing. Together, these frameworks are not abstract constructs, they are the cognitive tools experienced clinicians use daily to reason about patients as whole persons.

Yet a striking gap exists: while AI systems for healthcare have grown increasingly powerful, they have largely ignored the individualized patients characteristics grounded in psychological theory. Current applications of Large Language Models (LLMs) to clinical prediction rely either on generic prompting of feeding patient data without any structured reasoning guidance or on domain-specific fine-tuning, which may capture statistical patterns but might fails to reasons about patients in clinically meaningful ways [17,2]. Neither approach gives the model what a skilled clinician possesses: a *theory of the individual:* a structured, theoretically grounded account of how biological, psychological, and social factors interact within a specific individual to produce a specific pain response under specific conditions.

In this paper, we examine whether encoding validated psychological theory directly into LLM reasoning transforms clinical prediction from population-level pattern matching into individualized inference grounded in a structured understanding of the patient. We introduce *Psychology-Grounded Artificial Mental Models* a framework that encodes established psychological theories of individual differences directly into LLM reasoning, enabling it to construct an individualized mental model of each patient before predicting perceived pain on a (1–5) Numeric Rating Scale (NRS) during knee and lower-limb rehabilitation exercises. Given a structured patient profile and exercise context, the Biopsychosocial model and ICF framework are operationalized as a structured prompting architecture, instructing the LLM to integrate physical capacity, psychological disposition, lifestyle, and functional context into a holistic patient persona from which predictions are derived.

The distinction has direct predictive consequences. A generic LLM assigns a pain score to a 55-year old patient with high anxiety, poor sleep, and low emotional stability based on surface correlations. Our framework reasons over the patient persona: anxiety amplifies pain perception, poor sleep lowers pain thresholds,

and low emotional stability predicts catastrophizing, arriving at substantially more individualized prediction. This paper makes three contributions:

1. **A theoretically grounded reasoning framework for clinical prediction:** We demonstrate that established psychological theories of individual differences, including the Biopsychosocial model and the ICF framework, can be operationalized as structured LLM reasoning scaffolds, creating what we term *Psychology-Grounded Artificial Mental Models*. This bridges clinical psychology theory and AI-driven prediction.
2. **Systematic empirical evidence across LLMs and prompts:** Through experiments across four LLM families (LLaMA 3.1, Qwen 2.5, Mistral, and Phi 3) and multiple prompting strategies (zero-shot baseline, in-context learning, BPS-only, ICF-only, BPS+ICF, and exploratory pain-theory variants), we show that psychology-grounded prompting improves individualized prediction over generic baselines.
3. **Validation across two rehabilitation-related datasets:** We evaluate the framework on the DiscoverYourself study with 336 participants and further validate it on the publicly available PhysioPain dataset ($N = 82$ after pre-processing). The BPS+ICF framework achieves consistent improvements across model families, with gains of up to 56.9 percentage points in exact accuracy over the generic baseline on the DiscoverYourself dataset.

1.1 Background and Related Work

The Biopsychosocial (BPS) model [5] establishes that health outcomes emerge from the dynamic interplay of biological, psychological, and social factors rather than biological pathology alone. In rehabilitation, this matters profoundly: two patients with identical knee surgeries can follow entirely different recovery trajectories depending on fear of movement, personality-driven coping strategies, and available social support [14]. The ICF [24] complements the BPS model by providing a structured taxonomy of functioning across body functions, activities, participation, and contextual factors. Together, these frameworks give clinicians a principled vocabulary for reasoning about individual patients; the BPS model supplies the causal structure, the ICF supplies the descriptive completeness.

Johnson-Laird's theory of mental models [12] proposes that humans reason not through formal logic but by constructing internal representations of situations that capture the relationships between entities and their properties. Applied to clinical reasoning, an experienced clinician builds a mental model of each patient, an integrated, dynamic representation of who this person is, how they are likely to respond, and why. Our framework operationalizes this concept computationally: we construct Artificial Mental Models (AMMs) by encoding psychological theory as structured reasoning scaffolds within LLMs, enabling them to build an analogous patient representation before making predictions [10].

LLMs have demonstrated strong capabilities in clinical text analysis, medical question answering, and patient simulation [18,23,4]. Yet common paradigms for clinical prediction, such as generic prompting or domain-specific adaptation,

may fall short for individualized reasoning. Generic prompting provides no structured guidance for integrating heterogeneous patient factors. Domain-specific adaptation on small clinical datasets can risk capturing surface correlations rather than clinically meaningful patterns, and produces reasoning that is neither interpretable nor verifiable against clinical theory [2]. Our framework addresses this gap directly by encoding established psychological theory as explicit reasoning scaffolds, giving LLMs the same structured understanding of individual differences that clinicians rely on.

2 Psychology-Grounded AMM Framework

We formulate individualized pain prediction as an ordinal estimation task. Given a structured participant profile x_i and an exercise context e, the objective is to predict the participant's self-reported perceived difficulty $\hat{y}_i$ on a five-point Numeric Rating Scale (NRS):

$$f(x_i, e) = \hat{y}_i, \qquad \hat{y}_i \in \{1, 2, 3, 4, 5\} \tag{1}$$

where 1 denotes *not difficult at all* and 5 denotes *extremely difficult*. Critically, the target is a subjective self-report; the model must capture how an individual *experiences* an exercise, not merely whether they can physically perform it.

Our framework operationalizes this through three sequential components: (i) Psychology-driven input structuring, (ii) persona construction, and (iii) theory-guided prediction. Figure 1 illustrates the overall architecture.

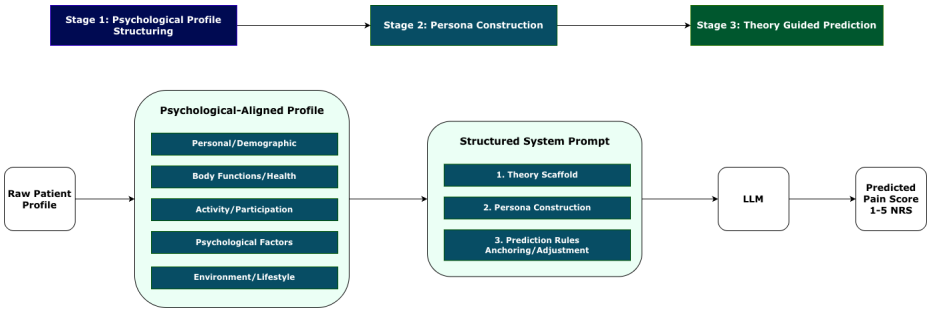

Fig. 1. Psychology-grounded AMM framework for pain-score prediction: Raw patient profiles are organized into theory-aligned biopsychosocial sections, converted into a structured system prompt with persona and prediction constraints, and passed to an LLM for scoring on a 1-5 NRS pain scale.

2.1 Psychology-Driven Input Structuring

A standard approach to clinical prediction treats patient attributes as a flat feature vector and assign all reasoning about interactions and context to the model. Our

framework instead imposes the structure of the BPS model and the ICF framework on the input itself, so that the LLM encounters patient information already organized according to established clinical reasoning categories. Concretely, each patient profile is partitioned into five theory-aligned sections that jointly cover the BPS and ICF dimensions:

1. **Personal / Demographic Factors** (ICF: Personal Factors): Age, gender, employment status, industry, work type.
2. **Body Functions / Health Condition** (BPS: Biological; ICF: Body Functions & Structures): Disability status, sleep hours, sleep problems, overall health, mobility, surgery history, surgical complications, recovery period, physical therapy history and adherence.
3. **Activity / Participation** (ICF: Activities & Participation): Daily activities performed without difficulty, ability to perform the target exercise (e.g., 10 squats or sit-to-stand), maximum repetitions achievable.
4. **Psychological Factors** (BPS: Psychological; ICF: Personal Factors): Current mood, stress or anxiety in the past 24 hours, emotional state, Big Five personality traits[4], daily-life stress level.
5. **Environmental / Lifestyle Factors** (BPS: Social; ICF: Environmental Factors): Types of physical activities practiced, exercise frequency (days per week), session duration, mood–exercise relationship (whether the participant exercises more when in a positive or negative mood).

2.2 Persona Construction

The system prompt instructs the LLM to process the structured patient profile based on mental model theory [12,9] and construct the internal representation persona of the patient. Rather than mapping features directly to a prediction, the model must first internally construct a realistic behavioral persona of the participant. The persona integrates eight dimensions:

- Physical capacity and mobility limitations
- Health conditions and recovery status
- Confidence or caution tendencies
- Current emotional state and stress exposure
- Exercise habits and physical activity level
- Personality-driven self-rating tendencies
- Whether the participant typically rates difficulty as low, moderate, or high
- Likely gap between perceived and actual difficulty

[4] Assessed via the Ten-Item Personality Inventory (TIPI) [7]: extraversion, agreeableness, conscientiousness, emotional stability, openness to experience, plus their reverse-scored counterparts.

2.3 Theory-Guided Prediction

The structured profile and the constructed persona together form the complete input to the LLM. Rather than presenting raw patient attributes directly to the model, the framework ensures that the model receives an psychological-interpreted representation. The prediction is derived from this representation, not from the raw features. The system prompt encodes the reasoning process in three stages. First, the model processes the theory-structured input and constructs the internal persona as described in Section 2.2. Second, guided by the encoded psychological frameworks, the model reasons over the persona to assess how the individual's biological, psychological, and social factors are likely to interact under the given exercise context. Third, the model produces a score on the target ordinal scale, where the prediction reflects the individual's expected subjective experience rather than an objective physical capacity estimate. This design ensures that prediction is grounded in a theoretically coherent account of the individual.

2.4 Prediction Mechanism

All LLM-based configurations use direct model generation followed by rule-based score extraction. Given a structured prompt p, the model generates a short textual response:

$$r = \mathrm{LLM}(p) \tag{2}$$

The generated response r is then parsed to obtain the predicted score $\hat{y} \in \{1, \ldots, 5\}$. The prompt explicitly instructs the model to output only one score from 1 to 5 and not to include reasoning. After generation, the response is normalized and checked for a valid ordinal score. If a valid score is found, it is used as the prediction; otherwise, the output is treated as invalid according to the evaluation protocol.

Using the same generation-and-parsing procedure across all configurations ensures that comparisons between prompting strategies reflect differences in the reasoning scaffold rather than differences in the prediction mechanism. All configurations therefore produce predictions on the same discrete 1–5 scale.

3 Data

We evaluate our framework on two independent datasets differing in provenance, population, and target outcome. Full dataset details are provided in Appendix C.

3.1 DiscoverYourself Dataset

The primary dataset comprises N=336 participant profiles collected across two waves of the DiscoverYourself study. In both waves, participants completed a structured digital questionnaire and performed ten bodyweight squats guided by video instruction. Participants self-reported perceived difficulty before and after

the exercise on a 1–5 Numeric Rating Scale (NRS), where 1 indicates *not difficult at all* and 5 indicates *extremely difficult* [16]. The post-exercise NRS score is the prediction target. The questionnaire captures variables spanning all BPS and ICF dimensions as described in Section 2.1: demographics, health and body functions, activity and participation, psychological state, and lifestyle and environmental factors. Variables are expressed as natural-language text rather than numeric codes, enabling the LLM to leverage the semantic content of responses during reasoning. The sample skews young and physically active, which we discuss as a limitation in Section 5.

3.2 PhysioPain Dataset

The PhysioPain dataset [21] comprises $N = 82$ participant profiles after preprocessing. The original survey contains a larger set of responses ($N = 99$), but we retain only records with a valid response for the target item, "Rate the severity of your pain (Likert scale)", corresponding to column AH in the source file. Records without this target value are omitted from evaluation. The dataset is a publicly available pain assessment survey covering demographics, health status, psychological state, and lifestyle factors. The prediction target is a self-reported pain outcome on the same 1–5 NRS scale. This dataset serves as an external validation set: collected through a different modality, from a partially different population, and targeting pain perception rather than exercise difficulty, it tests whether the framework generalizes beyond the primary study setting.

4 Experiments and Results

4.1 Experimental Setup

We evaluate four instruction-tuned LLM families: LLaMA 3.1 8B Instruct [8], Qwen 2.5 7B Instruct [19], Mistral 7B Instruct [11], and Phi-3 Mini Instruct [1]. Each model is evaluated under five primary configurations: a generic baseline, In-Context Learning (**ICL**), where 10 labeled examples for the DiscoverYourself dataset and 5 for the PhysioPain dataset are provided; **BPS-only**, where the system prompt encodes the BPS framework; **ICF-only**, where the prompt is structured around functional health categories; and **BPS+ICF**, the full framework with combined theoretical scaffolding, persona construction, and prediction rules. We additionally evaluate three exploratory pain-theory prompt variants: Fear-Avoidance, Cognitive-Behavioral, and Gate Control Theory prompting.

4.2 Results on DiscoverYourself Dataset

Table 1 reports exact accuracy across all model family on the DiscoverYourself dataset (N=336). The BPS+ICF configuration produces the strongest result for all four model families, with exact accuracy ranging from 62.8% to 63.7%. This consistency across LLaMA 3.1, Phi-3, Mistral, and Qwen 2.5 suggests that

Table 1. Exact accuracy (%) on the DiscoverYourself dataset (N=336). Best result per row in **bold**.

Model	Baseline	ICL(10-shots)	BPS-only	ICF-only	BPS+ICF
LLaMA 3.1 8B	7.4	19.6	54.8	29.5	**63.7**
Phi-3 Mini 4K	6.6	27.9	32.5	9.2	**63.5**
Mistral 7B v0.3	7.4	22.3	31.9	27.4	**62.8**
Qwen 2.5 7B	8.0	38.6	7.0	19.7	**63.4**

the gain is not tied to a single architecture. In particular, Qwen 2.5 improves substantially under the full BPS+ICF configuration, indicating that the combined theoretical scaffold is more stable than partial theory-only prompting for this dataset.

4.3 Results on PhysioPain Dataset

Table 2 reports exact accuracy on the PhysioPain dataset (N=82) targeting pain perception rather than exercise difficulty. The framework also generalizes

Table 2. Exact accuracy (%) on the external pain dataset (N=82). Best result per row in **bold**.

Model	Baseline	ICL(5-shots)	BPS-only	ICF-only	BPS+ICF
LLaMA 3.1 8B	39.0	35.0	25.6	25.7	**58.5**
Qwen 2.5 7B	31.7	37.6	50.0	23.6	**53.7**
Phi-3 Mini 4K	43.9	31.1	38.0	26.3	**58.5**
Mistral 7B v0.3	35.4	34.0	37.3	21.1	**54.9**

to the external PhysioPain dataset, with BPS+ICF achieving the strongest result for all four model families. LLaMA 3.1 and Phi-3 both reach 58.5% exact accuracy, followed by Mistral at 54.9% and Qwen 2.5 at 53.7%. Although the dataset is smaller and targets pain severity rather than post-exercise difficulty, the consistent performance of BPS+ICF suggests that the combined theoretical scaffold transfers beyond the primary 10-squat setting.

4.4 Ablation: Contribution of Psychological Theory

To quantify the contribution of psychological grounding, we compare the generic baseline against the full BPS+ICF framework on the DiscoverYourself dataset. The baseline represents prediction without the full psychology-grounded reasoning scaffold, while BPS+ICF adds theory-driven input structuring, persona construction, and combined biopsychosocial and functional reasoning. Figure 2 reports exact accuracy for the four base LLMs. Across all four model families, BPS+ICF

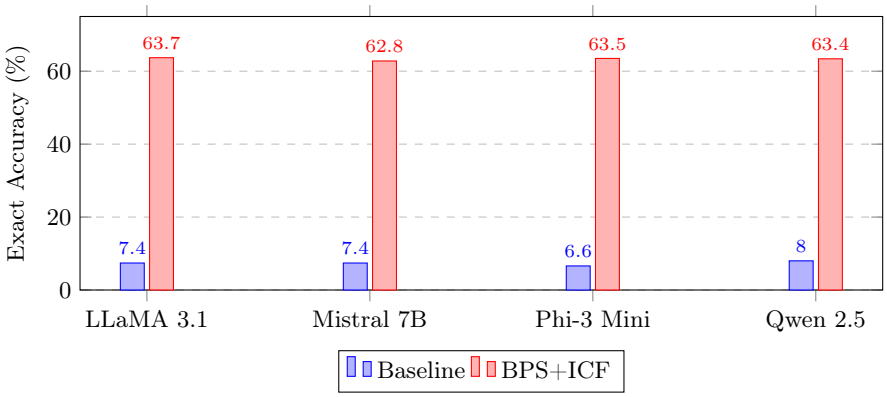

Fig. 2. Exact accuracy of the baseline and BPS+ICF framework on the DiscoverYourself dataset. BPS+ICF improves across all models, with gains of +55.4 to +56.9 percentage points.

substantially outperforms the baseline. LLaMA 3.1 improves from 7.4% to 63.7% exact accuracy (+56.3 pp), Mistral 7B improves from 7.4% to 62.8% (+55.4 pp), Phi-3 Mini improves from 6.6% to 63.5% (+56.9 pp), and Qwen 2.5 improves from 8.0% to 63.4% (+55.4 pp). These results show that the gain is consistent across model families and is not tied to a single architecture. The consistent improvement over the baseline indicates that theory-grounded prompting changes how the models use participant information. Rather than treating survey fields as isolated variables, the BPS+ICF framework encourages the model to reason over the interaction between biological condition, psychological state, lifestyle context, activity limitations, and participation-level factors. This supports the central claim that individualized rehabilitation prediction benefits from explicitly encoding psychological and functional theory into the LLM inference process.

4.5 Exploratory Pain-Theory Prompt Variants

To verify that the final BPS+ICF configuration was not selected arbitrarily, we additionally evaluated three pain-theory prompt variants: Fear-Avoidance, Cognitive-Behavioral, and Gate Control Theory prompting. These variants are treated as exploratory alternatives rather than as the main proposed framework. Unlike BPS+ICF, which combines biological, psychological, social, functional, participation-level, and contextual factors, these theories focus on narrower mechanisms of pain perception, avoidance, coping, and sensory modulation. Table 3 reports the results on the DiscoverYourself dataset, while Table 4 reports the corresponding results for the PhysioPain dataset.

Table 3. Exploratory exact accuracy (%) of pain-theory prompt variants on the DiscoverYourself dataset (N=336).

Model	Fear-Avoidance	Cognitive-Behavior	Gate Control
LLaMA 3.1 8B	4.5	3.9	4.5
Phi-3 Mini 4K	22.6	12.2	14.0
Mistral 7B v0.3	15.2	10.4	11.9
Qwen 2.5 7B	48.5	7.7	8.6

Table 4. Exploratory exact accuracy (%) of pain-theory prompt variants on the PhysioPain dataset (N=82).

Model	Fear-Avoidance	Cognitive-Behavior	Gate Control
LLaMA 3.1 8B	18.3	18.3	18.3
Phi-3 Mini 4K	31.5	53.7	56.1
Mistral 7B v0.3	23.2	30.5	25.6
Qwen 2.5 7B	50.0	43.9	32.9

5 Discussion

Why theory-grounded prompting improves prediction: The main finding is that BPS+ICF prompting consistently improves over the generic baseline across all four model families on the DiscoverYourself dataset. We attribute this improvement to the structure imposed by the psychological reasoning scaffold. Instead of treating participant variables as isolated fields, the prompt organizes them into clinically meaningful biological, psychological, social, functional, and contextual dimensions. This gives the model an explicit framework for interpreting heterogeneous participant information before producing a prediction.

Implications for aging and rehabilitation: Aging populations present two interrelated challenges: data scarcity and high individual heterogeneity [13]. Our framework addresses both by encoding established psychological theory into the prompt, providing domain knowledge that would otherwise require large-scale training data to approximate. Rather than learning from thousands of patient profiles that pain, reduced mobility, psychological state, and social context interact, the model is explicitly guided to reason through these relationships using BPS and ICF principles. Constructing an individualized persona per patient also makes the framework suitable for heterogeneous rehabilitation settings, where patients with similar diagnoses may differ substantially in psychological disposition, lifestyle, and functional capacity.

Limitations: The DiscoverYourself dataset skews young and physically active, limiting direct generalizability to elderly rehabilitation cohorts; targeted evaluation with older adults remains necessary. The PhysioPain dataset is also small after preprocessing, with 82 usable records, so external validation results should be interpreted cautiously. The 1–5 NRS target is subjective and coarse, and exact-match accuracy may not fully capture clinical utility. All evaluations are

cross-sectional; whether psychology-grounded reasoning maintains its advantage across longitudinal rehabilitation trajectories remains an open question. Finally, although BPS+ICF performs consistently across the evaluated model families, the exploratory pain-theory variants show that narrower theoretical prompts can behave differently across architectures and datasets.

6 Conclusion

We presented a framework that encodes established psychological theory as structured reasoning scaffolds within LLMs, enabling individualized prediction of self-reported exercise difficulty and pain-related outcomes. Evaluated across four LLM families and two datasets, the BPS+ICF framework produces consistent improvements over the generic baseline, with gains of up to 56.9 percentage points on the DiscoverYourself dataset. The central finding is that individualized rehabilitation prediction benefits not only from model scale or data availability, but also from theory: providing LLMs with clinically grounded frameworks such as the Biopsychosocial model and ICF improves how they interpret heterogeneous participant profiles. Future work will extend evaluation to older clinical cohorts, incorporate longitudinal rehabilitation trajectories, and examine additional psychological theories of pain and recovery.

Acknowledgments

This work was partially funded by the German Federal Ministry of Education and Research (BMBF) under the contract 01IW23004.

Ethics and Broader Impact Statement

Study Ethics and Data Collection. The DiscoverYourself dataset was collected under the oversight of the research institution's ethical review procedures. All participants provided informed consent prior to enrollment and were informed that their self-reported data would be used for AI model development. Participation was voluntary, with no clinical intervention administered and no therapeutic decisions influenced by the study. All participant records were pseudonymized prior to analysis, with identifying information removed from the dataset. The PhysioPain dataset [21] is publicly available under the CC BY 4.0 license [5].

Ethical Considerations of Psychological Profiling. A distinctive feature of our approach and one that requires explicit ethical reflection is that it encodes established psychological theories into an AI system's reasoning process, instructing LLMs to construct holistic behavioral personas of individual patients. While this is precisely what makes the approach effective, it also means that the system

[5] `https://creativecommons.org/licenses/by/4.0/`

engages in structured psychological profiling, integrating personality traits, emotional states, coping tendencies, and lifestyle patterns into an internal model of "who this patient is." We emphasize three safeguards. First, the constructed persona is an *internal reasoning scaffold*, not a stored or communicated artifact: it exists only within the model's inference process and is not saved, shared, or made accessible outside the prediction pipeline. Second, the reasoning is grounded in *established, peer-reviewed clinical frameworks* (BPS model, ICF) rather than ad-hoc or algorithmically discovered categories, ensuring that the profiling aligns with accepted clinical practice. Third, the persona is constructed from *self-reported* data that participants voluntarily provided, not from inferred or behavioral data collected without their knowledge. Despite these safeguards, we acknowledge that computational psychological profiling carries inherent risks. We strongly advocate that any deployment of psychology-grounded AMMs be subject to institutional ethics review, with clear governance over data access, model outputs, and downstream use.

Fairness and Representational Limitations. The DiscoverYourself dataset skews toward young, physically active participants, which limits the representativeness of our evaluation for the elderly populations most relevant to clinical deployment. The psychological frameworks we employ (BPS model, ICF, TIPI personality assessment) were developed and validated primarily in Western clinical and research contexts; their applicability across diverse cultural, linguistic, and socioeconomic backgrounds should not be assumed without further validation. We did not conduct formal fairness audits (e.g., across age, gender, or ethnicity subgroups) due to insufficient subgroup sample sizes for reliable comparison, and we flag this as a necessary step before any clinical deployment.

Reproducibility. All experiments use publicly available, instruction-tuned LLMs (LLaMA 3.1 8B, Qwen 2.5 7B, Mistral 7B v0.3, Phi-3 Mini 4K) accessed through standard inference APIs. The system prompts encoding the BPS and ICF frameworks are described in detail in Appendix A; exact prompt templates and code are available at [6].

References

1. Abdin, M., Aneja, J., Awadalla, H., et al.: Phi-3 technical report: A highly capable language model locally on your phone (2024), https://arxiv.org/abs/2404.14219
2. Barnett, S., Brannelly, Z., Kurniawan, S., Wong, S.: Fine-tuning or fine-failing? debunking performance myths in large language models. arXiv preprint arXiv:2406.11201 (2024)
3. Beck, A.T., Rush, A.J., Shaw, B.F., Emery, G., DeRubeis, R.J., Hollon, S.D.: Cognitive therapy of depression. Guilford Publications (2024)
4. Cao, W., Qu, M., Zhu, T., et al.: Benchmarking large language models against human experts in rehabilitation medicine: a multidimensional evaluation. Journal of

[6] GitHub:https://anonymous.4open.science/r/anon-repo-0661/README.md

NeuroEngineering and Rehabilitation **23**, 84 (2026). https://doi.org/10.1186/s12984-026-01903-0
5. Engel, G.L.: The need for a new medical model: a challenge for biomedicine. Science **196**(4286), 129–136 (1977). https://doi.org/10.1126/science.847460
6. Gatchel, R.J., Peng, Y.B., Peters, M.L., Fuchs, P.N., Turk, D.C.: The biopsychosocial approach to chronic pain: scientific advances and future directions. Psychological Bulletin **133**(4), 581–624 (2007). https://doi.org/10.1037/0033-2909.133.4.581
7. Gosling, S.D., Rentfrow, P.J., Swann Jr, W.B.: A very brief measure of the big-five personality domains. Journal of Research in personality **37**(6), 504–528 (2003)
8. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
9. Janzen, S., Saxena, P., Agnes, C., et al.: Ki in der rehabilitation – anwendung künstlicher mentaler modelle für eine personalisierte medizin. Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz **68**, 889–897 (2025). https://doi.org/10.1007/s00103-025-04090-w
10. Janzen, S., Maass, W., Saxena, P.: Investigation of artificial mental models for healthcare ai systems
11. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023), https://arxiv.org/abs/2310.06825
12. Johnson-Laird, P.N.: Mental models. (1989)
13. Khan, H.T.A., Addo, K.M., Findlay, H.: Public health challenges and responses to the growing ageing populations. Public Health Challenges **3**(3), e213 (2024). https://doi.org/10.1002/puh2.213
14. Kunnen, M., Dionigi, R.A., Litchfield, C., Moreland, A.: Psychological barriers negotiated by athletes returning to soccer (football) after anterior cruciate ligament reconstructive surgery. Annals of Leisure Research **26**(4), 545–566 (2023)
15. Leeuw, M., Goossens, M.E.J.B., Linton, S.J., Crombez, G., Boersma, K., Vlaeyen, J.W.S.: The fear-avoidance model of musculoskeletal pain: current state of scientific evidence. Journal of Behavioral Medicine **30**(1), 77–94 (2007). https://doi.org/10.1007/s10865-006-9085-0
16. Nugent, S.M., Lovejoy, T.I., Shull, S., Dobscha, S.K., Morasco, B.J.: Associations of pain numeric rating scale scores collected during usual care with research administered patient reported pain outcomes. Pain Medicine **22**(10), 2235–2241 (2021)
17. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Natarajan, V., et al.: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (2023). https://doi.org/10.1038/s41586-023-06291-2
18. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Natarajan, V., et al.: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (2023). https://doi.org/10.1038/s41586-023-06291-2
19. Team, Q.: Qwen2.5: A party of foundation models (September 2024), https://qwenlm.github.io/blog/qwen2.5/

20. Timm, A., Schmidt-Wilcke, T., Blenk, S., Studer, B.: Altered social decision making in patients with chronic pain. Psychological Medicine **53**(6), 2466–2475 (2023)
21. Toktay, B., İkbal Işık Orhan, Yıldırım, E., Akbulut, F.P., Catal, C.: Multimodal insights into diverse pain experiences: Physiopain dataset. Data in Brief **62**, 111992 (2025). https://doi.org/https://doi.org/10.1016/j.dib.2025.111992, https://www.sciencedirect.com/science/article/pii/S2352340925007164
22. Wade, D.T., Halligan, P.W.: The biopsychosocial model of illness: a model whose time has come. Clinical Rehabilitation **31**(8), 995–1004 (2017). https://doi.org/10.1177/0269215517709890
23. Wang, R., Milani, S., Chiu, J.C., Zhi, J., Eack, S.M., Labrum, T., Murphy, S.M., Jones, N., Hardy, K., Shen, H., et al.: Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. arXiv preprint arXiv:2405.19660 (2024)
24. World Health Organization: International Classification of Functioning, Disability and Health: ICF. World Health Organization, Geneva (2001), https://iris.who.int/handle/10665/42407

A Biopsychosocial–ICF Prompt Template

This appendix provides a condensed version of the BPS–ICF system prompt. Participant-specific survey fields were inserted separately during inference.

```
You are an expert in human behavior modeling and functional health interpretation.

Given participant survey data from an exercise study, predict the participant's
self-reported ACTUAL difficulty after performing 10 squats.

Internally construct a behavioral persona rather than treating fields as isolated
variables. Consider physical capacity, health and mobility limitations, emotional
state, stress or anxiety, exercise habits, confidence, and likely self-rating style.

Use perceived difficulty before exercise as the default prediction. Change the
score only when the full profile gives clear evidence that actual difficulty
would be higher or lower. If uncertain, stay close to perceived difficulty.

Use the combined Biopsychosocial and ICF frameworks internally:
- BPS: biological, psychological, and social/lifestyle factors.
- ICF: body functions, activities, participation, and contextual factors.

Return only one score from 1 to 5.
1 = not difficult at all; 2 = slightly difficult; 3 = moderately difficult;
4 = very difficult; 5 = extremely difficult.
```

B Theoretical Descriptions

Biopsychosocial model. The biopsychosocial model explains health and illness as interactions between biological, psychological, and social factors rather than as purely biomedical conditions [5]. In this study, it structures prompts around physical health, emotional state, confidence, lifestyle, and activity habits.

ICF. The International Classification of Functioning, Disability and Health describes functioning through body functions and structures, activities, participation, and contextual factors [24]. We use it to frame the 10-squat task as an activity-level outcome influenced by functional capacity and individual context.

C Dataset Descriptions

C.1 DiscoverYourself Dataset

The DiscoverYourself dataset contains 336 participant records from exercise assessments involving the 10-squat task. Although collected across multiple sessions or sources, the records are treated as one combined rehabilitation dataset. Each profile includes demographic, health, psychological, lifestyle, and task-specific variables. The prediction target is the participant's self-reported actual difficulty after performing 10 squats, represented on a 1–5 ordinal scale. Perceived difficulty before exercise is included as an input feature, while post-exercise actual difficulty is used as the ground-truth label.

C.2 PhysioPain Dataset

The PhysioPain dataset is used as an external validation dataset beyond the 10-squat setting. After excluding records without a valid target response, 82 participant profiles remain. The dataset contains demographic, health, pain-related, psychological, and lifestyle variables. The target is a self-reported pain-severity outcome mapped to a 1–5 ordinal scale. This tests whether the same Biopsychosocial–ICF prompting framework can transfer from exercise-difficulty prediction to pain-severity prediction.

D Implementation Details

All experiments were implemented in Python using PyTorch and the Hugging Face `transformers` library. Models were loaded with `AutoTokenizer` and `AutoModelForCausalLM`, using `bfloat16` precision and automatic device placement via `device_map="auto"`. If the tokenizer did not define a padding token, the end-of-sequence token was used as the padding token. The maximum input length was set to 4096 tokens.

Ground-truth and perceived-difficulty labels were parsed from the dataset using a rule-based extraction function. Exact accuracy was computed over records with a valid ground-truth label. Records without a valid target label were excluded from metric calculation.