Artificial Mental Modeling by Leveraging Prescriptive Components in Large Language Models

Motivation & Problem: Mental illness, neuro-degenerative diseases, medical rehabilitation applications critically depend on understanding individual cognitive patterns and emotional responses [Bickmore et al., 2022; Sharma et al., 2023]. However, current LLMs excel at population-level responses while fundamentally failing to model individual mental states and expectations. Existing personalization approaches rely on crude fine-tuning or prompt engineering that conflates general knowledge with individual patterns, leading to catastrophic forgetting and poor generalization. [Li et al., 2023; Kumar & Singh, 2024]. Recent work has shown that LLMs exhibit decision-making patterns comprising two distinct components, a descriptive component that reflects statistical norms from training data, and a prescriptive component that encodes implicit normative ideals, desirable, or values. [Bear & Knobe, 2017]. While prior research has focused on how these LLMs biases create problematic decision-making [Sivaprasad et al., 2025], we present a novel framework that leverages these prescriptive mechanisms that benefits individual mental modeling.

Approach: We propose Artificial Mental Modeling (AMM), a personalization framework for transformer models leveraging a key insight from cognitive science: individual mental models combine descriptive knowledge (statistical knowledge of the world) with prescriptive biases (personal normative expectations) [Johnson-Laird, 1983; Norman, 2014]. Our work extends the foundational approaches of [Sivaprasad et al., 2025; Zhu et al., 2024, Wang et al., 2024] by explicitly decomposing internal representations into a descriptive subspace that captures population-level knowledge and a prescriptive subspace that encodes individual-specific goals and priors $h = h_{desc}+h_{presc}$. Motivated by [Voita et al., 2019; Michel et al., 2019; Clark et al., 2019] that the transformer attention heads specialize and can be selectively pruned or repurposed, AMM partitions attention heads accordingly: “descriptive heads” are frozen to preserve general world knowledge, while “prescriptive heads” are adapted with parameter-efficient modules via adapters (LoRA) attached to the prescriptive heads [Hu et al., 2022]. We constrain the updates with orthogonality and sparsity regularizers to minimize interference with the descriptive subspace and mitigate catastrophic forgetting [Ravfogel et al., 2020; Kirkpatrick et al., 2017]. This separation enables targeted personalization without degrading performance across tasks requiring both generalization and individual tailoring.

Experiments & Results: We validated AMM framework on medical rehabilitation dataset [Janzen et al, 2025] comprising a corpus of web-scraped medical dialogues (n=4,364) and an empirical user study (n=116) with “patient profiles” encompassing five constructs: demographics, psychosocial factors, medical history, personality traits [TIPI; Gosling et al., 2003], and exercise-specific perceived difficulty/pain [Chiarovano et al., 2018; Goldman et al., 2014]. Participants rated expected pain (by watching a guided exercise video) and actual pain after performing the exercise on a (0–5) numerical rating scale (NRS). 40% of participants exhibited discrepancy between expected and actual pain (>2 NRS), indicating individual prescriptive biases our framework aims to capture. We therefore formalize the task as predicting individual post-exercise pain (NRS 0-5) across six architectures (GPT-4o-mini, LLaMA-2/3, Mistral, Phi-3) using zero-shot, few-shot, and fine-tuning approaches. AMM partitions attention heads into descriptive (frozen) and prescriptive (adapted) using a simple gradient*activation saliency score from [Kim & Lee 2025] computed on a small personalization set (we select the top 20% heads as prescriptive). Only prescriptive heads receive LoRA (rank = 64) adapters with orthogonality penalty limits update projections onto the descriptive subspace. Baselines include zero/few-shot prompting and fine-tuning (LoRA on all heads) on demographic/medical data. Results demonstrate AMM's effectiveness: LLaMA-3 8B achieved 92% individual pain prediction accuracy, outperforming descriptive only baselines by +42.9 percentage points (92% vs 49.1%). AMM shows consistent improvements across all architectures e.g. Mistral 7B gains +25.4 points (82% vs 56.6%), Phi-3 gains +40.3 points (79.9% vs 39.7%). Smaller models outperform larger ones under AMM: LLaMA-3 8B (92%) exceeds LLaMA-3 70B (31.87%), validating our hypothesis that large models’ descriptive subspaces can dominate sparse prescriptive signals during adaptation. Overall, AMM consistently outperformed descriptive-only baselines for capturing individual patterns across all model sizes proving empirical demonstration that prescriptive biases can benefit individual modeling.

Future Work: Our experiments are limited by sample size, self-reported pain, heuristic head selection; future work will replace heuristic head selection with a learned mask trained under sparsity and orthogonality constraints. We plan to overcome current data/profile limitations by validating at scale on the National Health and Nutrition Examination Survey (NHANES) dataset and conducting a rigorous safety/fairness audit on RAI benchmarks (RealToxicityPrompts, HarmBench, Toxic/WildChat etc).

GitHub: https://github.com/ArtificialMentalModel/AMMs