Offline Policy Evaluation of Multi-Turn LLM Health Coaching with Real Users

Melik Ozolcer **Sang Won Bae**
{mozolcer, sbae4}@stevens.edu
Stevens Institute of Technology

Abstract

We study a web-deployed, tool-augmented LLM health coach with real users. In a pilot with seven users (280 rated turns), offline policy evaluation (OPE) over factorized decision heads (TOOL/STYLE) shows that a uniform heavy-tool policy raises average value on logs but harms specific subgroups, most notably low-health-literacy/high-self-efficacy users. A lightweight simulator with hidden archetypes further shows that adding a small early information-gain bonus reliably shortens trait identification and improves goal success and pass@3. Together, these early findings indicate an evaluation-first path to personalization: freeze the generator, learn subgroup-aware decision heads on typed rewards (objective tool outcomes and satisfaction), and always report per-archetype metrics to surface subgroup harms that averages obscure.

1 Introduction

Wearable devices generate personal health data that AI health coaches powered by LLMs can analyze in real time. These agents can help users understand trends, assess risks, and achieve health goals. However, successful deployment requires more than accurate analysis: agents must maintain trust across multi-turn dialogues, manage long conversations, personalize to individual health profiles, and handle sensitive data competently. Current research relies mainly on synthetic benchmarks, leaving gaps in understanding real-world performance with actual users.

This paper presents early findings from a web-deployed LLM health agent and outlines an operational framework for personalization. We (i) perform offline policy evaluation (OPE) on deployment logs to compare decision policies (e.g., when to use tools, response style), and (ii) run a lightweight simulation with hidden user archetypes to test whether brief early exploration improves outcomes. Analysis of pilot interactions shows an early period of high-rated, tool-driven engagement that degrades as conversation length increases. While baseline health literacy and self-efficacy are associated with user experience, tool usage and response characteristics emerge as the primary drivers of rating variation. Both OPE and simulation reveal heterogeneous effects across archetypes.

We model health coaching as a user-conditioned partially observable Markov decision process (POMDP) with typed rewards. The per-turn utility is a user-specific composition defined in Eq. 1, where z_t summarizes dialogue context, recent outcomes, and user characteristics (including health literacy and self-efficacy), and the tool component derives from success/failure outcomes. To address cold-start and sparse feedback, we add a small, early-turn user-model information-gain bonus (curiosity) that encourages actions which reduce uncertainty over latent interaction preferences (e.g., tolerance for explanation, desire for analysis) [1–4]. We define this bonus formally in Eq. 2 and apply it only in the first few turns.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Multi-Turn Interactions in Large Language Models.

First, OPE with self-normalized IPS (SNIPS) and AIPW (doubly robust) [5–7] lets us compare counterfactual policies on real logs without additional user trials. We find that uniform, heavy tool usage can maximize average reward on logs but causes harm for specific archetypes (e.g., low-literacy/high-efficacy), strengthening the case for archetype-aware reward weighting. Second, in simulation with hidden archetypes, brief early information-gain phase reliably shortens trait-identification time and improves task success and reliability (pass@3), which supports a "probe-then-personalize" strategy consistent with our deployment findings.

2 Related Work

Health LLMs and tool-using agents. LLM-based agents have advanced health applications in coaching and diagnostics. PH-LLM [8] fine-tunes models on wearable data for personalized sleep and fitness insights, while PHIA [9] uses iterative reasoning with tools for health data interpretation across multiple turns. Recent work explores behavior-theoretic scaffolding [10] and proactive strategies [11] for long-term coaching. However, multi-turn deployment remains challenging: [12] report substantial degradation over extended conversations, and tool-enabled settings can be brittle. Notably, [13] show that while tools and natural-language feedback generally help performance, multi-turn pipelines remain fragile and certain training choices can hurt longer conversations.

Personalization, curiosity, and reward modeling. Personalization through RL and preference learning shows promise for diverse users. Curiosity and intrinsic-motivation approaches use information gain to reduce uncertainty about user traits. These have been adapted to multi-turn dialogue, education, and fitness, where early exploration can accelerate downstream performance [1]. Related work includes variational preference learning with latent user variables [14], turn-level credit assignment for tool-using agents [15], and low-rank reward modeling for efficient personalization [16]. We build on these ideas but focus on health coaching with tool use outcomes and survey-informed archetypes (health literacy, self-efficacy). We combine an early information-gain bonus with user-conditioned reward composition.

Evaluation: OPE, simulators, and long-horizon benchmarks. Off-policy evaluation (OPE) lets us compare counterfactual policies from logs without needing new user trials, using importance-weighted and doubly-robust estimators with calibration diagnostics [5–7]. Tool-using agent benchmarks now use multi-turn tasks with typed end-state checks and pass@3 reliability [13, 17, 18], while long-horizon assistants expose planning and consistency limitations [19, 20]. We align with this trend by (i) applying OPE on deployment logs to compare decision policies under selection correction and (ii) using a lightweight simulator with hidden archetypes to test whether early information-gain phases improve trait identification, goal success, and reliability.

3 Methods

3.1 System and Data

We built and deployed a web-based LLM health system (Qwen3-235B-A22B via API providers [21]) for multi-turn health coaching with real users and wearable data. The system processes user-uploaded health data (i.e., Apple Health app export) through: (1) a data pipeline that preprocesses historical data into daily features (sleep, HRV, VO_2max, activity) and generates ML predictions for stress, soreness, and injury risk (rolling-origin cross-validation; $R^2 = 0.50,\ 0.28,\ 0.40$ respectively on 1,280 day-samples, $N=28$) [22]; (2) agent tools including *Code Executor*, *Web Searcher*, and *Email* (reminders and events).

Participants and Procedure. Seven students ($N=7$) completed a pilot study involving baseline surveys (eHEALS for health literacy [23] and the General Self-Efficacy Scale [24]), followed by multi-turn dialogues over multiple days (typically >1 hour, $30+$ messages) with per-response 1-5 star ratings. We logged system metrics (latency, token counts, tool usage/outcomes) and client-side interaction proxies (typing duration, edits/pauses/tab-switches). The study received IRB approval and participants provided informed consent.

3.2 Tool-Outcome Rubric

We labeled each tool invocation using a rubric that drives the objective component of reward R_{tool}:

Table 1: Typed tool-outcome rubric and mapping to R_{tool}. We define R_{tool} on every turn; when no tool is invoked, R_{tool}=0.

Tool	Outcome	Criteria	R_{tool}
No Tool	—	no tool invoked	0
Web Searcher	**Success**	$\geq$1 URL; claim-evidence consistent	$+1$
	Failure	no URL or dead/contradictory link	-1
Code Executor	**Success**	executes; plot/table present; axes/units labeled	$+1$
	Failure	exception; wrong data slice; no visualization	-1
Email	**Success**	event/message created; ISO8601 date/time present	$+1$
	Failure	not created / confirmation missing	-1

Note: R_{eng} is a small, bounded engagement signal defined on *all* turns (e.g., mild latency penalty, structure/clarity bonus, evidence gating). The OPE objective uses $R_{\text{obj}} = R_{\text{tool}} + R_{\text{eng}}$ and is reported via SNIPS.

3.3 Problem Formulation

We model multi-turn coaching as a user-conditioned partially observable Markov decision process (POMDP). At turn t, the agent receives observation o_t (user latest tool outputs) and maintains a belief state $z_t = f_\phi(h_t, u_i, m_t)$ summarizing conversation history h_t, user features u_i (eHEALS/GSE strata and derived interaction features), and current metrics m_t (health data, model predictions, recent tool outcomes). The action a_t factorizes into discrete heads:

$$a_t = (\underbrace{\quad\quad}_{\text{TOOL}\ \in\{\varnothing,\ \text{Search},\ \text{Code},\ \text{Email}\}}, \underbrace{\quad\quad}_{\text{STYLE}\ \in\{\text{concise},\ \text{detailed}\}})$$

with stochastic policy $\pi_\theta(a_t \mid z_t, u_i)$ producing per-head probabilities.

3.4 Reward Signals and Personalization

Per-turn utility is a personalized composition

$$R_i(z_t, a_t) \;=\; \alpha_i(z_t)\, R_{\text{user}}(z_t, a_t) \;+\; \beta_i(z_t)\, R_{\text{tool}}(\tau_t) \;+\; \gamma_i(z_t)\, R_{\text{eng}}(z_t, a_t), \tag{1}$$

where $\tau_t \in \{\text{success}, \text{failure}\}$ is the logged tool-outcome label (Table 1), and $R_{\text{tool}}(\varnothing) \equiv 0$ when no tool is invoked. All components are computed from observables in z_t and the chosen action a_t; we do not assume access to latent state.

Early-turn information-gain (curiosity). To mitigate cold-start and selection-biased feedback, we add a small, early-turn user-model information-gain bonus that rewards reductions in archetype uncertainty:

$$r_t^{\text{curiosity}} \;=\; \max\big\{0,\ H\big(p_{t-1}(y)\big) - H\big(p_t(y)\big)\big\}, \tag{2}$$

where y is a latent interaction archetype (literacy $\times$ self-efficacy) and $p_t(y)$ is the posterior of a lightweight user model combining dialogue features and survey priors. The intrinsic weight λ_t is applied only in the first K turns (we use $\lambda \in \{0.1, 0.2\}$; K=2 in our simulator) and then decays to zero. The total per-turn signal for learning/evaluation is $R_i + \lambda_t\, r_t^{\text{curiosity}}$, with curiosity complementing (never replacing) R_{user} and R_{tool}.

3.5 Offline Policy Evaluation (OPE)

We evaluate counterfactual decision policies over (TOOL, STYLE) from deployment logs, following logged-bandit practice [5–7]. We interpret OPE as a per-turn contextual bandit evaluation conditioned on z_t, and do not claim unbiased long-horizon policy value.

Table 2: Off-Policy Evaluation Diagnostics on Pilot Logs

Diagnostic	Value	Notes
Data Quality		
Sessions	23	82.6% agreement between `conv_len` and `turn_index`; 1.1% boundary conflicts
Rating rate	0.80	Fraction of turns with ratings
OPE Correction		
Rating-propensity AUC	0.712	For AIPW importance weighting
Clipping rate ($c = 50$)	$\leq$0.29%	Low importance weight explosion
Policy Calibration (ECE)		
TOOL head	0.157	Expected calibration error of behavior policy per prediction head
STYLE head	0.050	
Tool Performance		
Web search	81.6%	Success rates by tool type. Top failures: `no_url_provided`, `no_visualization`
Code execution	80.7%	
Email	85.7%	

Behavior-policy reconstruction and calibration. For each decision head we fit a probabilistic behavior model $\hat{\pi}_b^h(a^h \mid x_t)$ on features $x_t = (z_t, u_i)$ (turn features and user attributes) to approximate logging propensities. We assess calibration via Expected Calibration Error (ECE) per head. For primary reporting we use self-normalized importance sampling (SNIPS)[25] for the objective component and augmented inverse propensity weighting (AIPW) for satisfaction.

Behavior policy and importance ratios. We model the logging policy as head-factorized over $\{\textsc{Tool}, \textsc{Style}\}$ and fit cross-fitted, calibrated classifiers for each head to obtain per-head propensities $\hat{\pi}_b^h(a_t^h \mid x_t)$. Assuming head-wise factorization conditional on x_t, the joint logging propensity is $\hat{\pi}_b(a_t \mid x_t) = \prod_h \hat{\pi}_b^h(a_t^h \mid x_t)$, and the target policy factorizes analogously. For both the SNIPS estimator and the augmentation term in the AIPW estimator, we use the same joint importance ratio

$$w_t \;=\; \prod_{h \in \{\textsc{Tool},\textsc{Style}\}} \frac{\pi^h(a_t^h \mid x_t)}{\hat{\pi}_b^h(a_t^h \mid x_t)}\,,$$

with ratio clipping at c=50 and session-level bootstrap confidence intervals.

Estimators and clipping. We clip all importance ratios at c=50. Our primary estimators are:

$$\widehat{V}_{\text{SNIPS}}(\pi) = \frac{\sum_t w_t\, r_t}{\sum_t w_t}, \qquad \widehat{V}^{\text{user}}_{\text{AIPW}}(\pi) = \frac{1}{T}\sum_t \left(\hat{q}_{\text{user}}(x_t, \pi) + \frac{m_t}{\hat{p}_{\text{rate},t}}\, w_t \left[r_t^{\text{user}} - \hat{q}_{\text{user}}(x_t, a_t) \right] \right),$$

where $m_t \in \{0, 1\}$ indicates whether a rating is observed and $\hat{p}_{\text{rate},t}$ is a rating-propensity model.

Total reward composition. For interpretability we report an unweighted objective $R_{\text{obj}} = R_{\text{tool}} + R_{\text{eng}}$ with SNIPS, alongside a R_{total} that applies fixed, pre-specified literacy-aware weights to form $R_{\text{total}} = \alpha_w R_{\text{user}} + \beta_w R_{\text{tool}} + \gamma_w R_{\text{eng}}$. Here $(\alpha_w, \beta_w, \gamma_w)$ is a deterministic function of literacy strata (low-literacy: $(0.6, 0.2, 0.2)$; high-literacy: $(0.3, 0.5, 0.2)$). We estimate $\mathbb{E}_\pi[\beta_w R_{\text{tool}} + \gamma_w R_{\text{eng}}]$ with SNIPS and $\mathbb{E}_\pi[\alpha_w R_{\text{user}}]$ with AIPW using a rating-propensity correction. Because R_{obj} (unweighted) and R_{total} (weighted) use different compositions and estimators (SNIPS vs. AIPW), columns are reported on their native scales and do not sum algebraically in Table 5.

Diagnostics. Calibration, selection, and coverage diagnostics (rating missingness and propensity AUC, per-head ECE, clipping rate) are reported in Table 2.

3.6 Simulator with Hidden Archetypes

To study early exploration and personalization without on-policy trials, we build a lightweight simulator with latent archetypes $y \in \{\text{L}_{\text{low/high}} \times \text{E}_{\text{low/high}}\}$. Each episode samples (i) a health

timeseries table (sleep/HRV/steps, etc.) and tasks requiring rolling computations/plots and (ii) wellness APIs (e.g., `set_reminder`, `log_sleep`) with verifiable end-states. The user-simulator returns a natural reply, a rubric-based rating, and tool outcome labels (Table 1). We compare policies with/without the curiosity term (Eq. 2) on: final return, goal success, pass@3, trait-identification turn/accuracy (from $p_t(y)$), and archetype-aligned action rate. In all simulator runs, the base generator is frozen; policies operate only through the discrete heads.

3.7 Reproducibility

We release de-identified logs of the rated conversational messages, and the code for all evaluations and simulations at `https://github.com/stevenshci/NeurIPS-MTI-LLM`.

4 Pilot Study Results

We analyzed 350 conversation turns with 280 ratings from our initial pilot with 7 users. While this sample size limits statistical conclusions, we report descriptive patterns that informed our framework design and motivated the new OPE and simulation studies. Ratings showed no meaningful temporal autocorrelation (lag-1: $r{=}0.10$).

4.1 User Baseline Influence

Users with above-median self-efficacy averaged $4.3{\pm}0.5$ ratings vs. $3.9{\pm}0.7$ for below-median. Table 3 shows baseline metrics' influence on interaction behaviors. User archetypes revealed varied satisfaction patterns: low literacy/high efficacy users reported highest median ratings (4.5), while high literacy/low efficacy reported lowest (3.8).

Table 3: Correlations between baseline metrics and interaction behavior ($N{=}7$). Values are descriptive.

Baseline Metric	Correlation (r)
Health literacy vs. tool usage	0.302
Self-efficacy vs. conversation length	-0.441

4.2 Tool Use Impact

Tool usage showed descriptive differences: tool-invoked responses averaged 4.02 vs. 4.38 for non-tool responses (Cohen's $d{=}-0.40$). Table 4 details category-specific patterns, with a success–failure gap of $+0.50$. Tool responses showed $1.40\times$ higher variance ($\sigma^2{=}0.94$ vs. 0.67), reflecting high-risk/high-reward dynamics.

Table 4: Impact of tool use on ratings (1-5) by category. (Email is n=7 turns, 1 failure).

Tool Category	Success Mean	Failure Mean	Gap
Web Searcher	4.01	3.53	+0.48
Code Executor	4.09	3.33	+0.76
Email	5.00	5.00	0.00
All Tools	4.08	3.58	+0.50

Mixed-effects modeling suggested minimal between-user heterogeneity in unconditional ratings: only 1.6% of variance was attributable to user identity (ICC=0.016). This does not preclude *context-conditioned* heterogeneity: as shown below, per-archetype differences under counterfactual policies are large.

4.3 Implicit Feedback and Performance Degradation

Response time negatively correlated with ratings ($r{=}-0.18$): $<$ 10s responses averaged 4.25 vs. 3.98 for $>$ 30s. Message length showed a small positive association ($r{=}0.07$); high-rated messages

Table 5: Offline policy evaluation on deployment logs. R_{obj}: SNIPS (unweighted objective); R_{user}: AIPW (z-score). R_{total}: literacy-weighted combination. Brackets show session-level bootstrap 95% CIs for R_{total}.

Policy	R_{obj} (SNIPS)	R_{user} (AIPW)	R_{total}
NoTool	0.328	-0.623	**0.044** [-0.045, 0.198]
AlwaysTool	0.229	-0.654	**0.304** [**0.001**, 0.524]
HeuristicGated	0.309	-0.625	**0.006** [-0.111, 0.174]
PersonalizedWeights	0.253	-0.656	**0.113** [-0.016, 0.284]

Table 6: Heterogeneity by archetype: difference in OPE estimates (AlwaysTool − NoTool). Positive is better.

Archetype	Δ Objective	Δ Satisfaction
$\text{L}_{\text{high}} \times \text{E}_{\text{high}}$	+0.575	-0.107
$\text{L}_{\text{high}} \times \text{E}_{\text{low}}$	+0.595	**+0.525**
$\text{L}_{\text{low}} \times \text{E}_{\text{low}}$	+0.165	-0.431
$\text{L}_{\text{low}} \times \text{E}_{\text{high}}$	-0.315	-1.436

(4–5) averaged 2,123 characters vs. 1,275 for low-rated (1–2). Ratings declined from 4.36 (turns 1–5) to 4.12 (15+), while tool usage peaked at 70% (turns 5–10) then dropped to 26.3% for turns 15+. Latency peaked mid-conversation (25s at turns 6–10), and message length decreased 3.4% from early to late turns.

4.4 Counterfactual Policy Analysis

Using SNIPS and AIPW with session-level bootstrap CIs, we compared four policies over discrete decision heads: NoTool (never invoke a tool), AlwaysTool (always invoke a tool; tool type drawn from the logged conditional distribution to maintain overlap), HeuristicGated (simple rule-based triggers), and PersonalizedWeights(same heuristics with a literacy-aware bias that raises or lowers the gate). On average, AlwaysTool attained the highest R_{total} on our logs (0.304, 95% CI [0.001, 0.524]). However, per-archetype slices revealed large and structured heterogeneity: heavy tool use benefited high-literacy/low-efficacy users across both objective and satisfaction components, but harmed low-literacy/high-efficacy users on both (e.g., ΔObjective -0.315; ΔSatisfaction -1.436 vs. NoTool). Tables 5–6 summarize results.

4.5 Simulation with Curiosity

We built a lightweight simulator with hidden archetypes (literacy × self-efficacy) and verifiable tool tasks. Adding a small early information-gain bonus for the first two turns improved goal success (0.935 → 0.970), pass@3 (0.95 → 0.98), final return (-3.162 → -2.329), and reduced trait-identification time (6.41 → ≈5.7 turns), with a small expected dip in archetype-aligned actions due to early probing (Table 7).

Summary. The pilot study suggests three findings: (i) tool use is *high-variance* with typed, fixable failure modes; (ii) average trends mask *large, structured heterogeneity* by archetype; and (iii) performance degrades over turns, consistent with a need for targeted early information gathering and personalization. The OPE and simulator results build on these observations and provide preliminary validation for our evaluation-first, personalization-first framework.

5 Implications for Future RL Frameworks

We outline design implications and a blueprint for RL-style personalization informed by our OPE and simulator findings. We do not train an end-to-end RL policy in this work, and the base generator remains frozen. OPE indicates that a uniform heavy-tool policy can lift averages while harming specific subgroups (Table 6), and simulation suggests a brief early information-gain phase improves

Table 7: Simulation with hidden archetypes; λ controls the early information-gain bonus (applied in the first two turns). Trait-ID turn: lower is better.

Policy	Final Return	Goal Succ.	pass@3	Trait-ID Turn	Trait Acc.	Arch. Align
HEURISTIC	-2.908	0.515	0.505	6.315	0.050	0.503
PERSONALIZED	-3.162	**0.935**	**0.950**	6.415	0.050	0.424
PERS+CURIOSITY (λ=0.10)	-2.401	**0.965**	**0.975**	**5.655**	0.075	0.412
PERS+CURIOSITY (λ=0.20)	-2.329	**0.970**	**0.980**	5.860	0.045	0.410

HEURISTIC (mirror of HEURISTICGATED; STYLE defaults to *concise* unless the user asks to "explain/why/how"), PERSONALIZED (mirror of PERSONALIZEDWEIGHTS with STYLE also adapted by archetype).

task success and shortens trait identification (Table 7). Motivated by these observations, we frame coaching as a user-conditioned POMDP with a belief state $z_t = f_\phi(h_t, u_i, m_t)$ and a policy acting over discrete decision heads (TOOL/STYLE). As a hypothesis for future work, we propose archetype-aware weighting that increases β for $\mathrm{L_{high}} \times \mathrm{E_{low}}$ users, reduces β and raises α for $\mathrm{L_{low}} \times \mathrm{E_{high}}$, and pairs tools with explanatory scaffolds for $\mathrm{L_{low}} \times \mathrm{E_{low}}$; a short, bounded curiosity term (Eq. 2) can accelerate preference elicitation in the first turns.

Moreover, we recommend an offline-first loop: fit or select population head policies from logs via contextual-bandit objectives while keeping the generator frozen; initialize user-specific weights from cluster priors (eHEALS/GSE strata and early interaction features) and adapt with strong regularization; and gate deployments by OPE diagnostics. End-to-end policy learning and prospective validation are left to future work.

6 Conclusion

We studied a deployed, tool-augmented LLM health coach with real users and extended our observations with two components: (i) offline policy evaluation (OPE) on deployment logs to compare counterfactual decision policies, and (ii) a lightweight simulator with hidden archetypes to test brief early information-gain phases. Three findings emerge. First, while uniform heavy tool usage maximizes average reward on logs, OPE reveals large heterogeneity across archetypes: the same policy benefits high-literacy/low-efficacy users but harms low-literacy/high-efficacy users on both objective and satisfaction measures. Second, adding a small, bounded early information-gain bonus consistently improves simulated task success, pass@3, and reduces trait-identification time, supporting a "probe-then-personalize" strategy. Third, the OPE pipeline (behavior-policy reconstruction, calibration diagnostics, clipping, session-bootstrap CIs, and rating-selection correction) is stable enough-per diagnostics, to guide offline-first policy iteration without new user trials.

These results provide an operational path forward: freeze the generator, learn subgroup-aware decision heads (instantiated here as literacy-aware) using typed rewards (objective outcomes + satisfaction), and use OPE to compare policies offline. Critically, report per-archetype metrics to catch policies that improve average performance while harming specific user groups.

Limitations and next steps. Our deployment sample is small and demographically narrow, limiting generalizability. Behavior propensities are reconstructed rather than logged, and Tool-head calibration is moderate, which can bias IPS/SNIPS; we mitigate with AIPW, clipping, and diagnostics (ECE). Our current evaluation weights and personalized routing are conditioned on literacy only; incorporating self-efficacy into weighting and head policies is a planned extension.

References

[1] Yanming Wan, Jiaxing Wu, Marwa Abdulhai, Lior Shani, and Natasha Jaques. Enhancing personalized multi-turn dialogue with curiosity reward. *arXiv preprint arXiv:2504.03206*, 2025.

[2] Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991.

[3] Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2): 265–286, 2007.

[4] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

[5] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International conference on machine learning*, pages 814–823. PMLR, 2015.

[6] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

[7] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International conference on machine learning*, pages 2139–2148. PMLR, 2016.

[8] Justin Cosentino, Anastasiya Belyaeva, Xin Liu, Nicholas A Furlotte, Zhun Yang, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, et al. Towards a personal health large language model. *arXiv preprint arXiv:2406.06474*, 2024.

[9] Mike A Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, et al. Transforming wearable data into health insights using large language model agents. *arXiv preprint arXiv:2406.06464*, 2024.

[10] Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James A Landay. Gptcoach: Towards llm-based physical activity coaching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–46, 2025.

[11] Vidya Srinivas, Xuhai Xu, Xin Liu, Kumar Ayush, Isaac Galatzer-Levy, Shwetak Patel, Daniel McDuff, and Tim Althoff. Substance over style: Evaluating proactive conversational coaching agents. *arXiv preprint arXiv:2503.19328*, 2025.

[12] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.

[13] Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv:2309.10691*, 2023.

[14] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. *Advances in Neural Information Processing Systems*, 37:52516–52544, 2024.

[15] Siliang Zeng, Quan Wei, William Brown, Oana Frunza, Yuriy Nevmyvaka, and Mingyi Hong. Reinforcing multi-turn reasoning in llm agents via turn-level credit assignment. *arXiv preprint arXiv:2505.11821*, 2025.

[16] Avinandan Bose, Zhihan Xiong, Yuejie Chi, Simon Shaolei Du, Lin Xiao, and Maryam Fazel. Lore: Personalizing llms via low-rank reward modeling. *arXiv preprint arXiv:2504.14439*, 2025.

[17] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. tau-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

[18] Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks. *arXiv preprint arXiv:2503.15478*, 2025.

[19] Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2023.

[20] Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. Travelplanner: A benchmark for real-world planning with language agents. *arXiv preprint arXiv:2402.01622*, 2024.

[21] Qwen Team. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

[22] Melik Ozolcer and Sang Won Bae. Sepa: A search-enhanced predictive agent for personalized health coaching. In *IEEE-EMBS International Conference on Biomedical and Health Informatics 2025*, 2025.

[23] Cameron D Norman and Harvey A Skinner. eheals: the ehealth literacy scale. *Journal of medical Internet research*, 8(4):e507, 2006.

[24] Ralf Schwarzer and Matthias Jerusalem. Generalized self-efficacy scale. *J. Weinman, S. Wright, & M. Johnston, Measures in health psychology: A user's portfolio. Causal and control beliefs*, 35(37):82–003, 1995.

[25] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28, 2015.