PWM-Pilot-Audio — Advisor / Reviewer Summary

One-page technical overview for academic feedback

Yuri Sylvester

June 6, 2026

PWM-Pilot-Audio — Advisor / Reviewer Summary

Status: Pre-pilot protocol draft — pre-registration candidate — no empirical results.

Premise

Personal-AI "understanding" is almost always evaluated retrospectively (surveys, recall, preference reconstruction), which a fluent-but-unaccountable system can pass. PWM (Person World Model) reframes the measurement: **understanding is the capability; forecasting is the operational measurement.** A system that genuinely models a specific person should produce *calibrated, person-specific* forecasts of that person's near-future behavior that beat (1) a population prior and (2) the person's own routine. Forecasting is treated as a falsifiable proxy for understanding, not as identical to it. This sits alongside prospective benchmarks like ForecastBench, but forecasts a *specific individual's* behavior against that individual's own baseline.

Design

Prospective, longitudinal, within-participant, evidence-tier ablation. **5 adults × 30 days**, daily forecasts sealed (timestamped + SHA-256 hashed) before outcomes. Audio-first. **Unit of analysis = the resolved forecast, not the participant.** Explicitly **not** powered for population inference; the goal is instrument validation (scoring correctness, calibration behavior, evidence-tier separability, resolution reliability, feasibility).

Evidence tiers (ablation ladder)

R1 population prior · R2 personal routine (strict walk-forward) · L0 metadata · L1 text · **L2 audio transcript (primary novel tier)**. L3 video excluded, reserved for future work. Central contrasts: **L2 vs L0/L1** and **L2 vs R2**.

Tasks & endpoints

Six task families (next contact, event realization, response behavior, attention allocation, routine deviation, commitment follow-through), each with a frozen answer space and resolution method. **Primary endpoint:** Personal Skill vs R2 on **attention allocation (T4)**. **Secondary:** Personal Skill vs R2 on **commitment follow-through (T6)**.

Metric & integrity

- **Proper scoring rules:** logarithmic (primary), Brier (secondary).
- **Personal Skill (bits)** = (mean log-score of system − mean log-score of reference) / ln 2, reported vs R1 and vs R2.
- **Calibration:** ECE / reliability diagrams / Brier decomposition (pass <=0.10, warn 0.10-0.20, fail >0.20), thresholds fixed pre-analysis.
- **Identity-permutation test:** score person i's forecasts against person j's outcomes; genuine person-specific skill must collapse toward zero (true skill > 95th pct of permuted null).
- **Leakage control:** hard 20:00 cutoff + sealing + hash verification; any post-cutoff timestamp in forecast inputs = automatic violation.
- **Inference:** paired-by-forecast comparisons, day-blocked bootstrap, leave-one-participant-out sensitivity. No generalization claims.

The two questions a reviewer should press

1. **Is R2 strong enough?** If L2 only beats a naive routine model, the result is meaningless. R2 uses recency-frequency + exponential decay + weekday/time-of-day + task-specific estimation; hyperparameters frozen pre-analysis.
2. **Does skill survive permutation?** If it does, the system is exploiting population structure, not the individual — and the "person-specific" claim fails.

Bounded claims

No claim that audio captures the full person; no claim forecasting equals understanding; no clinical/affective inference (acoustic features only as labeled exploratory metadata); no medical, diagnostic, or therapeutic claims. A pilot where nothing beats R2, or skill survives permutation, is a **successful, informative** pilot that localizes what to fix before scaling.

What remains before running

IRB/legal review per jurisdiction · implement & validate R1/R2 · unit-test scoring/cutoff/sealing/permutation on synthetic data · freeze per-participant T4 topic sets and prompt versions · end-to-end dry run · complete pre-registration.

Feedback most wanted on: baseline strength, permutation pass-criterion, answer-space resolvability, and whether the instrument-validation framing is honestly scoped.