PWM-Pilot-Audio

A Prospective Evaluation of Person-Specific Forecasting from Longitudinal Audio Evidence

Yuri Sylvester

June 6, 2026

Contents

1 Title Page

PWM-Pilot-Audio

A Prospective Evaluation of Person-Specific Forecasting from Longitudinal Audio Evidence

Status	Pilot protocol draft — pre-registration candidate — no empirical results yet
Prepared by	Yuri Sylvester
Date	June 6, 2026
Document type	Pre-pilot protocol
Version	0.1 (draft for review)

Note. This protocol is intended for academic feedback, participant review, and possible pre-registration before data collection. It describes a study that has **not** yet been run. It contains no empirical results, makes no claims of clinical or medical benefit, and does not assert that audio recording captures a person's true mental state. All quantitative statements about expected performance are hypotheses or design targets, not findings.

2 Executive Summary

Personal AI systems increasingly claim to "understand" their users. In practice, that claim is almost always evaluated *retrospectively* — through satisfaction surveys, self-report agreement, memory recall, or preference reconstruction. None of these establish that a system can be held accountable for statements about a person's *future*. A system that paraphrases your past well may still be useless at anticipating what you will actually do tomorrow.

This protocol operationalizes a different evaluation philosophy, which we call PWM (Person World Model). The governing idea is simple: **understanding is the capability; forecasting is the measurement.** If a system genuinely models a specific person, that model should produce *calibrated, person-specific forecasts* of that person's near-future behavior that beat two strong, honest baselines: (1) a population prior that knows nothing about the individual, and (2) the individual's own routine. Forecasting here is not a claim that prediction equals understanding — it is an operational, falsifiable measurement of progress toward it.

PWM-Pilot-Audio is a small, prospective, within-participant pilot: **5 consenting adults, 30 days, daily sealed forecasts**, using **audio-derived evidence** as the primary novel signal. Audio is chosen as the first rich evidence tier because speech externalizes goals, plans, commitments, and concerns; because it is substantially cheaper and less invasive than video; and because it lets us validate the entire scoring and integrity machinery before any multimodal expansion. Video (L3) is explicitly out of scope here and reserved for future work.

The **primary scientific question** is whether longitudinal audio-derived evidence improves person-specific forecasting *beyond* population priors and personal routine. The **primary forecasting targets** are **attention allocation** (which project/topic gets the most attention tomorrow) and **commitment follow-through** (whether a stated commitment is acted on within 72 hours). The **primary metric** is **Personal Skill in bits** — the per-forecast improvement in logarithmic score over a reference model, expressed in bits — reported separately against the population prior (R1) and the personal routine baseline (R2).

Two integrity tests guard against the most likely ways to be fooled. **Calibration analysis** (expected calibration error, reliability diagrams, Brier decomposition) ensures a system cannot "win" through overconfident lucky guesses. The **identity-permutation test** re-scores a participant's forecasts against *other* participants' outcomes; genuine person-specific skill should collapse toward zero under permutation, whereas skill that survives permutation indicates the system is exploiting population structure rather than modeling the individual.

The **privacy posture** is conservative and explicit. Raw audio is the most sensitive asset in the study; it remains private to the participant and researcher, is processed local-first wherever practical, is never released, and is never published. Only derived evidence, aggregate scores, and synthetic examples leave the secure environment. Participants may pause recording at any time and may flag any segment for deletion.

The pilot is deliberately **not powered for population-level claims.** Its unit of analysis is the *resolved forecast*, not the participant. Its purpose is to validate the protocol, the scoring methodology, the evidence-tier comparison, and operational feasibility — and to determine whether a larger, IRB-supervised benchmark round is warranted. A pilot in which no system beats R2, or in which skill fails to collapse under permutation, is still a successful pilot: it tells us precisely where the benchmark, the evidence pipeline, or the task definitions must be repaired before scaling.

3 Plain-Language Participant Summary

This section is written for potential participants. It avoids technical terms. A standalone one-page version is provided as Appendix A-handout.

What is this? This is a research study about whether a computer system can predict a person's near-future everyday choices — like which project they'll focus on tomorrow, or whether they'll follow through on something they said they'd do. We want to find out whether recordings of your everyday speech help the system make better predictions than simply guessing from averages or from your usual routine.

What would I actually do? For 30 days you would record portions of your day's audio using a recorder or phone app, and spend about 5–10 minutes each evening confirming a few things that happened (for example, "Did the meeting happen?" or "Which project did you spend the most time on?"). That's the core of it.

How much time does it take? The recording mostly runs in the background while you go about your day. The active work is the short daily review — roughly 5 to 10 minutes. Initial setup takes about 30–45 minutes once.

What does "recording audio" mean here? It means capturing sound from your everyday environment — mostly your own speech and conversations you're part of. You decide when it runs. You can pause it any time, and you can delete anything you don't want kept.

What will be shared, and what will not? Your **raw audio will never be shared or published.** It stays private to you and the researcher, stored encrypted. From the audio we create text transcripts and short structured notes (for example, "mentioned a deadline Friday"). Only **summarized, anonymized numbers** — how accurate the predictions were on average — are ever used in any report. Your name is never attached to files; a code like "P001" is used instead.

Is it voluntary? Can I stop? Yes, completely voluntary. You can stop at any time, for any reason, with no penalty, and you can ask for your data to be deleted.

A few important cautions. Audio is sensitive. Conversations may accidentally capture **other people who haven't agreed to be recorded.** Please do not record where other people reasonably expect privacy, and please minimize capturing others; if it happens, you can flag that segment for deletion. Recording laws also differ by location — we'll go over what applies to you.

What this is not. This is **research**, not medical care, therapy, diagnosis, or productivity coaching. The system is not judging you, treating you, or giving you advice. Nothing here is a health assessment. We are studying prediction methods, not you as a patient.

If anything here is unclear, ask before agreeing. You should only take part if you understand and are comfortable with all of the above.

4 Scientific Motivation

4.1 The retrospective-evaluation gap

Most claims that an AI system "understands" a user are validated by methods that look backward. Satisfaction surveys ask whether output felt right. Self-report studies ask whether a summary matched the person's self-image. Recall and preference-prediction tasks ask whether a model can reproduce already-known facts. Personalization metrics measure engagement with content the system already surfaced. Each of these can be satisfied by a system that is fluent and agreeable without being *accountable*: none of them commits the system, in advance, to a falsifiable statement about what the person will do next.

This matters because retrospective agreement is cheap to fake and easy to overfit. A model can be highly persuasive about your past while being no better than chance about your future. If we want a measurement that resists this failure mode, the measurement must be **prospective**: forecasts must be sealed *before* outcomes occur, and then scored against what actually happened.

4.2 Forecasting as the measurement of understanding

The PWM framing inverts the usual order. Rather than asking a model to describe a person and then judging the description, we ask the model to *forecast* the person and then judge the forecast with a proper scoring rule. The premise is not that prediction is identical to understanding — a barometer predicts rain without understanding weather. The premise is weaker and defensible: **a system that genuinely models a specific person should be able to out-forecast both a population prior and that person's own routine, and should do so with calibrated probabilities.** Forecasting is therefore treated as an *operational proxy* — a necessary, measurable signature of person-specific modeling — not as a definition of understanding.

This places PWM-Pilot-Audio in the lineage of prospective forecasting benchmarks such as ForecastBench, which scores models only on questions whose answers are unknown at submission time, while differing in a critical respect: the quantity being forecast is not a public world event but a *specific individual's* near-future behavior, evaluated against that individual's own baseline.

4.3 Why audio is a reasonable first evidence tier

Audio is the natural first rich modality for four reasons. First, **speech externalizes the relevant latent variables**: people say what they intend to do, what they are worried about, what they have committed to, and what they are paying attention to. These future-oriented utterances are exactly the raw material a person-specific forecaster needs. Second, **audio is cheaper and less invasive than video** — lower storage, simpler capture, and far less sensitive than continuous visual recording of a person's environment. Third, **audio is sufficient to validate the full protocol** — capture, evidence extraction, sealing, resolution, scoring, calibration, and permutation — before committing to the cost and privacy burden of multimodal data. Fourth, audio **degrades gracefully**: even imperfect transcripts retain most of the future-oriented content that matters for forecasting.

4.4 Language discipline

We are deliberately careful about what is and is not claimed. Audio does **not** capture "the full person"; it captures a partial, self-selected slice of externalized speech. Forecasting skill does **not** equal understanding; it is a measurable correlate. We do **not** infer mood, affect, or clinical state

from acoustic features; any such signal is treated only as exploratory metadata, never as a diagnostic label (see §13). Throughout, "skill" means improvement over a named reference under a proper scoring rule — nothing more.

5 Research Questions

RQ1 — Beating the population prior. Can evidence-based systems forecast a person's future attention, interactions, events, responses, and commitments better than a population prior that has no person-specific information?

RQ2 — Beating personal routine. Can evidence-based systems beat the person's own routine baseline (their historical behavior)?

RQ3 — Marginal value of audio. Does audio transcript evidence (L2) add predictive skill beyond calendar/task/message metadata (L0) and text evidence (L1)?

RQ4 — Person-specificity. Does person-specific skill collapse when participant identity is permuted (i.e., when forecasts for person i are scored against outcomes of person j)?

RQ5 — Task forecastability. Which task families are most forecastable in a small pilot, and which are dominated by noise or ambiguous resolution?

RQ6 — Participant burden. How burdensome is the audio-first protocol in practice — recording load, daily-review time, and dropout risk?

RQ7 — Definitional adequacy. Are the outcome definitions and answer spaces clear and resolvable enough to support a larger benchmark round without post-hoc changes?

6 Hypotheses

Hypotheses are stated directionally and are to be fixed before any data are examined. They are pilot-scale expectations, not powered predictions.

H1. Evidence-based systems (L0–L2) will achieve **positive Personal Skill over the R1 population prior**.

H2. At least one evidence-based system will achieve **positive Personal Skill over the R2 personal routine baseline**.

H3. Audio transcript evidence (L2) will improve Personal Skill over metadata/text-only evidence (L0/L1), especially for **attention allocation (T4)** and **commitment follow-through (T6)**.

H4. Person-specific skill will **collapse toward zero under identity permutation**.

H5. Models will perform **better on near-term concrete events** (e.g., event realization) **than on longer-horizon goal-state transitions** (e.g., multi-day commitment follow-through and routine deviation).

Null hypothesis (H0). No system beats R2 on the primary endpoints. Failure to reject H0 is an acceptable and informative pilot outcome.

7 Study Design

PWM-Pilot-Audio is a **prospective, longitudinal, within-participant, evidence-tier ablation** study. Its defining methodological commitment is that **every forecast is sealed before the corresponding outcome occurs**, eliminating hindsight and post-hoc rationalization.

Design attribute	Specification
Orientation	Prospective (forecasts precede outcomes)
Time structure	Longitudinal, 30 consecutive days per participant
Comparison	Within-participant; each participant is their own control
Manipulation	Evidence-tier ablation (R1, R2, L0, L1, L2)
Sealing	Forecasts timestamped and cryptographically hashed before outcomes
Sample	5 consenting adults
Cadence	Daily forecasts; next-day and 72-hour horizons
Evidence	Audio-first (up to L2); video/L3 excluded
Unit of analysis	The **resolved forecast**, not the participant

The within-participant ablation is what gives a 5-person study analytical traction: across 30 days and six task families, each participant contributes hundreds of resolved forecasts, and the scientific comparisons (L2 vs L0/L1, evidence vs R2) are made *within* each person's own stream of outcomes. This isolates the marginal value of evidence tiers from inter-person heterogeneity.

What the pilot is not. It is not powered for population-level inference. We will make **no generalization claims** about humans in general, about demographic subgroups, or about effect sizes in a broader population. The pilot validates the *benchmark instrument*: protocol soundness, scoring correctness, calibration behavior, evidence-tier separability, resolution reliability, and operational feasibility.

8 Participants

Target sample. 5 consenting adults.

Eligibility (all required).

- Age 18 or older.
- Able to give informed consent.
- Comfortable recording portions of daily-life audio.
- Owns or can reliably use a recording device (dedicated recorder, smartphone, or wearable).
- Willing to complete a short daily review for 30 days.
- Understands and accepts the privacy risks described in §24.
- Able to pause or stop recording when needed.

Exclusion (any one disqualifies).

- Unwillingness to record audio.
- Inability to provide informed consent.
- Routine presence in high-risk or recording-prohibited environments (e.g., secured facilities, certain clinical or legal settings).
- Frequent, unavoidable capture of non-consenting third parties where mitigation is impractical.
- Legal or workplace restrictions that prohibit recording in the participant's normal environment.

Optional target profile (not a requirement, not a clinical criterion). Adults with high project load — founders, operators, knowledge workers, remote workers, or people who self-describe as having executive-function challenges — are an informative population because their attention allocation and commitments are varied and consequential. **This is not a medical study.** Recruiting people who describe executive-function challenges does **not** make this clinical research and confers no diagnostic or therapeutic intent. Any framing that touches clinical populations would require an academic or clinical partner and IRB oversight before proceeding.

9 Data Collection

Daily collection comprises up to four streams. Only stream A (audio) and stream C (daily review) are required; B and D are optional and consent-gated.

9.1 A. Audio recordings (required)

- **Target volume:** 8–12 hours/day of consenting-context audio where feasible.
- **Minimum acceptable:** 4 hours/day. Days below this threshold are flagged as low-coverage and handled per the analysis plan (§22).
- Participant may **pause at any time**.
- **No recording** in sensitive or prohibited settings.
- **No recording of private third-party conversations** without consent where applicable.

9.2 B. Calendar/task metadata (optional)

- Optional import or manual daily export of calendar and task data.
- Event titles may be included **only** if the participant consents; otherwise events are reduced to time/duration/category.

9.3 C. Daily participant review (required)

- **5–10 minutes/day.**
- Confirm major events of the day.
- Confirm which stated commitments were completed or not completed.
- Confirm attention allocation among predefined categories.
- Flag sensitive segments for deletion or exclusion.

9.4 D. Optional text evidence (optional)

- Notes, tasks, messages, or documents — **only** if explicitly consented.
- Not required for the audio-only pilot; used to populate the L1 tier when available.

9.5 What is explicitly not collected

- No medical diagnosis or health assessment.
- No covert recording.
- No secret or background monitoring outside participant-controlled capture.
- No intervention, nudge, or feedback to the participant during the scoring window (the system never tells the participant what it predicted before the outcome resolves).

10 Evidence Tiers

The evidence tiers form a strict inclusion ladder used for ablation. Each higher tier adds a category of evidence; baselines R1 and R2 use no contemporaneous personal evidence.

Tier	Name	Contents	Used in pilot?
R1	Population prior	Aggregate frequencies only; no personal evidence	Yes (baseline)
R2	Personal routine baseline	Participant's own historical routine only (walk-forward)	Yes (baseline)
L0	Metadata	Calendar / task / communication metadata	Yes
L1	Text evidence	Notes, tasks, messages, documents (if consented)	Yes (where available)
L2	Audio transcript evidence	Transcript, diarization, topics, entities, commitments, future-oriented statements	Yes (primary novel tier)
L3	Video / multimodal	Visual evidence	**No** — reserved for future study

Scope. PWM-Pilot-Audio tests up to **L2**. The central ablation contrast is **L2 vs. L0/L1** (does audio add skill beyond metadata and text?) and **L2 vs. R2** (does audio beat routine?). A future multimodal study will test whether **L3 adds marginal predictive skill beyond L2**; nothing in this pilot should be read as evaluating video.

11 Forecasting Tasks

Six task families are defined. Each has a fixed question, a fixed answer space, and a fixed resolution method. **Answer spaces are frozen before data collection** and may not be changed post hoc (§17).

11.0.1 T1 — Next Contact / Interaction

- **Question:** Who, or what category of person/entity, will the participant interact with next (or within a defined window)?
- **Answer space:** {family, work, friend, service provider, unknown/other, no interaction}.
- **Resolution:** message/call/calendar/audio evidence, or participant confirmation.

11.0.2 T2 — Event Realization

- **Question:** Will a specific planned event occur, cancel, move, be missed, or remain unclear?
- **Answer space:** {occurred, cancelled, rescheduled, missed, unresolved/excluded}.
- **Resolution:** calendar state plus participant confirmation.

11.0.3 T3 — Response Behavior

- **Question:** Will the participant respond to a specific message/request within a time window?
- **Answer space:** {no reply, <1 hour, 1–6 hours, 6–24 hours, >24 hours}.
- **Resolution:** message timestamps or participant confirmation.

11.0.4 T4 — Attention Allocation *(primary endpoint)*

- **Question:** Which active project/topic will receive the most attention in the next day?
- **Answer space (per-participant, fixed at enrollment):** e.g., {PWM-Bench, hardware, family, work/business, health/admin, other}.
- **Resolution:** transcript topic share, calendar/task activity, and participant daily review (adjudicated; see §17).

11.0.5 T5 — Routine Deviation

- **Question:** Will the participant substantially deviate from their expected routine tomorrow?
- **Answer space:** {no deviation, work deviation, family deviation, health deviation, travel/location deviation, social deviation, other}.
- **Resolution:** daily review plus evidence.

11.0.6 T6 — Commitment Follow-Through *(co-primary/secondary endpoint)*

- **Question:** A commitment was stated today. Will it be acted upon within 72 hours?
- **Answer space:** {completed, partially completed, not completed, superseded/no longer relevant, unresolved}.
- **Resolution:** observable action, document/task/message evidence, or participant confirmation.

11.0.7 Endpoint roles

Role	Task
Primary	T4 — Attention allocation
Co-primary / secondary	T6 — Commitment follow-through
Support (volume + protocol validation)	T1, T2, T3
Exploratory	T5 — Routine deviation

12 Daily Procedure

A single day proceeds as a fixed cycle. Times are local to the participant.

Phase	Time	Action
Morning	Start of day	Participant starts audio recording.
During day	—	Participant records normal consenting contexts; pauses as needed.
Evidence cutoff	**20:00 (8 PM)**	Hard cutoff. Only data timestamped at or before cutoff may inform forecasts.
Extraction	After cutoff	System extracts evidence from data up to cutoff **only**.
Forecast generation	After extraction	All systems generate forecasts for next-day / 72-hour tasks.
Sealing	Before any outcome	Forecasts are timestamped, **hashed**, and stored before outcomes can occur.
Next day	As events occur	Outcomes resolve as evidence becomes available.
Daily review	Evening (next day)	Participant confirms key outcomes and flags exclusions.
Weekly	Once/week	Researcher checks data completeness, participant burden, and protocol issues.

The cutoff/sealing boundary is the integrity backbone of the study: it makes future leakage detectable (any evidence ID with a timestamp after cutoff appearing in a forecast is an automatic protocol violation) and makes the forecast immutable once sealed (the stored hash can be re-verified at scoring time).

13 Audio Processing Pipeline

The pipeline is **local-first wherever practical** to minimize exposure of raw audio. Cloud components, if any, must be explicitly consented and must never receive raw audio that has not been reduced to consented derived evidence.

Suggested stack (all components replaceable):

Stage	Suggested tool(s)	Notes
Audio capture	Plaud, Apple Watch + Just Press Record, smartphone recorder, or equivalent	Participant-controlled start/pause
File storage	Encrypted local folder or NAS	No cloud for raw audio by default
Transcription	WhisperX or faster-whisper	Local inference preferred
Diarization	pyannote.audio or WhisperX diarization	Separate speakers
Speaker labels	participant / other / unknown	Where feasible
Segmenting	Split by day and time window	Aligns to cutoff boundaries
Entity extraction	Local NLP or LLM	People, orgs, dates, tasks
Commitment extraction	LLM prompt over transcript chunks	See Appendix G
Topic modeling	Embeddings + clustering, or LLM summaries	Drives T4 topic share
Evidence database	JSONL + SQLite/Postgres	Human-readable + queryable
Scoring	pwm-bench scoring harness	Proper scoring rules + permutation

Design notes consistent with maintainability: every stage writes a human-readable artifact (JSONL/SQLite) so the pipeline is observable and debuggable; each component is replaceable without rewriting the others; raw audio, transcripts, and derived evidence are kept in separate directories so storage and processing responsibilities stay clean.

> **Caution — no clinical inference from acoustics.** Do **not** infer mood, emotion, or any clinical state from audio. Acoustic features (e.g., speech energy, pace) may be retained only as **exploratory "affective cues" or "speech-energy proxies,"** clearly labeled as non-diagnostic metadata. They must never be used as outcome labels, diagnostic signals, or health indicators.

14 Evidence Extraction Schema

Each unit of derived evidence is one JSON record. Records are append-only; corrections are made by adding new records, never by silently editing history.

```
{
  "evidence_id": "P001-2026-06-06-0042",
  "participant_id": "P001",
  "timestamp_start": "2026-06-06T14:03:11-04:00",
  "timestamp_end": "2026-06-06T14:03:39-04:00",
  "source_type": "audio_transcript",
  "evidence_tier": "L2",
  "modality": "audio",
  "transcript_excerpt": "I need to send the grant draft to Dana before Monday.",
  "speaker": "participant",
  "evidence_type": "commitment",
  "entities": ["Dana", "grant draft"],
  "topics": ["work/business", "PWM-Bench"],
  "candidate_goal": "submit grant draft",
  "commitment": "send grant draft to Dana",
  "due_time": "2026-06-08T23:59:00-04:00",
  "attention_target": "work/business",
  "confidence": 0.82,
  "sensitivity_level": "normal",
  "excluded": false,
  "exclusion_reason": null,
  "model_used": "whisperx-large-v3 + extractor-v0.1",
  "prompt_version": "extract-commitment-v0.1"
}
```

Field reference.

Field	Meaning
`evidence_id`	Stable unique ID (participant + date + counter).
`participant_id`	Pseudonymous ID (never a name).
`timestamp_start` / `timestamp_end`	Evidence time window (used for cutoff enforcement).
`source_type`	e.g., audio_transcript, calendar, message, note.
`evidence_tier`	R1/R2/L0/L1/L2.
`modality`	audio / text / metadata.
`transcript_excerpt`	Minimal supporting quote (sensitive; stays private).
`speaker`	participant / other / unknown.
`evidence_type`	See enumerated types below.
`entities`	Named people/orgs/objects.
`topics`	Mapped to participant's fixed topic set.

Field	Meaning
`candidate_goal`	Inferred goal, if any.
`commitment`	Stated commitment, if any.
`due_time`	Commitment deadline, if stated/inferable.
`attention_target`	Topic this evidence indexes for T4.
`confidence`	Extractor confidence [0,1].
`sensitivity_level`	normal / sensitive / restricted.
`excluded`	Whether excluded from analysis.
`exclusion_reason`	Why excluded (audited).
`model_used`	Pipeline component + version.
`prompt_version`	Versioned prompt identifier.

Enumerated `evidence_type` values: commitment, decision, future_intention, concern, repeated_topic, attention_signal, routine_signal, social_interaction, event_update, uncertainty_signal.

15 Forecast Generation Procedure

Forecast generation is governed by five non-negotiable rules:

1. **No future evidence.** Each model receives only evidence with timestamps at or before the cutoff.
2. **Probabilistic output.** Each model outputs a probability distribution over the task's exact answer space; probabilities must sum to 1.
3. **Fixed answer spaces.** Forecasts must reproduce the frozen answer space exactly — no added, dropped, or renamed options.
4. **Sealed before outcomes.** Forecasts are hashed and stored before any outcome can resolve.
5. **Shared definitions.** All systems use identical task definitions, answer spaces, and cutoffs.

Systems under evaluation.

ID	System	Evidence available
A	R1 population prior	None (aggregate frequencies)
B	R2 personal routine baseline	Participant history only (walk-forward)
C	L0 metadata model	Calendar/task/communication metadata
D	L1 text model	L0 + text evidence
E	L2 audio transcript model	L0 + L1 + audio transcript evidence
F *(optional)*	Human self-prediction	Participant's own forecast

System F (the participant predicting their own next day) is an optional but valuable extra baseline: a person-specific model that cannot beat the *person's own self-prediction* is a meaningful negative result.

16 Forecast Prompts

All prompts below are versioned and reproduced in Appendix F. They are written to be **conservative**: every forecasting prompt instructs the model to output calibrated probabilities, to use only supplied pre-cutoff evidence, to avoid inventing private facts, to cite evidence IDs, to express uncertainty, and to preserve the answer space exactly.

16.1 Prompt A — System instruction (all forecasting models)

```
You are participating in PWM-Pilot-Audio, a prospective forecasting study.
You are given structured evidence about a single participant, identified only
by a pseudonymous ID, up to a fixed cutoff time t. Your task is to forecast a
future outcome for that participant.

Rules:
- Use ONLY the supplied evidence. Do not assume facts not present in it.
- Do NOT use any knowledge of events after the cutoff time t.
- Do NOT invent private facts about the participant.
- Output a CALIBRATED probability distribution over the EXACT answer space given.
- Probabilities must be non-negative and sum to 1.0.
- Cite the evidence_id(s) that most informed your forecast.
- Express genuine uncertainty; do not be overconfident.
- Output ONLY the requested JSON. Do not add commentary.
```

16.2 Prompt B — T4 Attention allocation

```
TASK: T4 - Attention allocation.
QUESTION: Which active project/topic will receive the MOST attention from the
participant during the next calendar day (00:00-23:59 local, the day after cutoff)?
ANSWER SPACE (use exactly these labels): {PWM-Bench, hardware, family,
work/business, health/admin, other}.
EVIDENCE (<= cutoff t): {evidence_records}
Return JSON:
{ "task":"T4", "participant_id":"...", "cutoff":"...",
  "distribution": { "PWM-Bench":0.0, "hardware":0.0, "family":0.0,
  "work/business":0.0, "health/admin":0.0, "other":0.0 },
  "cited_evidence": ["..."], "rationale_short":"<= 30 words" }
```

16.3 Prompt C — T6 Commitment follow-through

```
TASK: T6 - Commitment follow-through.
CONTEXT: The following commitment was stated by the participant on/before cutoff t:
"{commitment_text}" (evidence_id: {commitment_evidence_id}).
QUESTION: Will this commitment be acted upon within 72 hours of cutoff?
ANSWER SPACE: {completed, partially completed, not completed,
superseded/no longer relevant, unresolved}.
EVIDENCE (<= cutoff t): {evidence_records}
Return JSON:
{ "task":"T6", "participant_id":"...", "commitment_evidence_id":"...",
```

```
  "cutoff":"...",
  "distribution": { "completed":0.0, "partially completed":0.0,
  "not completed":0.0, "superseded/no longer relevant":0.0, "unresolved":0.0 },
  "cited_evidence": ["..."], "rationale_short":"<= 30 words" }
```

16.4 Prompt D — T2 Event realization

```
TASK: T2 - Event realization.
CONTEXT: A planned event is on record: "{event_text}" scheduled for {event_time}.
QUESTION: What will happen to this event?
ANSWER SPACE: {occurred, cancelled, rescheduled, missed, unresolved/excluded}.
EVIDENCE (<= cutoff t): {evidence_records}
Return JSON:
{ "task":"T2", "participant_id":"...", "event_ref":"...", "cutoff":"...",
  "distribution": { "occurred":0.0, "cancelled":0.0, "rescheduled":0.0,
  "missed":0.0, "unresolved/excluded":0.0 },
  "cited_evidence": ["..."], "rationale_short":"<= 30 words" }
```

16.5 Prompt E — Calibration instruction (appended to all forecasts)

```
Calibration requirement: your stated probabilities should match long-run
frequencies. If you say 0.70, then across many such forecasts the outcome should
occur about 70% of the time. Avoid assigning 0.0 or 1.0 unless the outcome is
logically impossible or certain given the evidence. Reserve extreme probabilities
for genuinely extreme evidence.
```

16.6 Prompt F — Evidence extraction (commitments/goals)

```
You extract structured, factual evidence from a transcript chunk. Do not infer
mental/clinical state. Do not fabricate. For each commitment, decision, future
intention, concern, or attention signal, output one record matching the evidence
schema. Use only what is explicitly or near-explicitly stated. Mark uncertain
items with lower confidence. Output JSON list only.
```

16.7 Prompt G — Daily summary

```
Summarize the participant's day from pre-cutoff evidence only, in <= 150 words.
Neutral, factual, non-clinical. List: stated commitments, planned events,
dominant topics, and any explicit future intentions. Cite evidence_ids.
Do not predict; only summarize what was said/recorded.
```

16.8 Prompt H — Outcome adjudication

```
You adjudicate the OBSERVED outcome of a forecasting task using post-cutoff
evidence and participant confirmation. Choose exactly one label from the task's
answer space, or "unresolved/excluded" if evidence is insufficient or the answer
space does not fit. Do not change the answer space. Cite the evidence_ids and the
resolution_method. If the participant disputes the resolution, set
adjudicator="participant-dispute" and excluded=true pending review.
```

17 Outcome Resolution

Outcomes are resolved with a strict, auditable procedure that prioritizes objective evidence and forbids post-hoc manipulation of the answer space.

Resolution principles.

- Objective evidence (timestamps, calendar state, observable action) is preferred.
- Participant confirmation is allowed when objective evidence is unavailable.
- Ambiguous outcomes are **excluded** or marked **unresolved** — never guessed.
- All exclusions are **logged before scoring**, with reasons.
- **No post-hoc answer-space changes.** If the answer space was wrong, the item is excluded, not retrofitted.

Outcome record schema.


```
{
  "forecast_id": "P001-T4-2026-06-06",
  "participant_id": "P001",
  "task_id": "T4",
  "resolution_time": "2026-06-08T09:15:00-04:00",
  "observed_answer": "PWM-Bench",
  "resolution_method": "transcript_topic_share+daily_review",
  "evidence_ids": ["P001-2026-06-07-0007", "P001-2026-06-07-0019"],
  "adjudicator": "researcher",
  "confidence": 0.9,
  "excluded": false,
  "exclusion_reason": null
}
```


Human adjudication rules.

1. If an outcome cannot be resolved reliably, **exclude** it.
2. If the frozen answer space does not fit the observed reality, **exclude** it (do not invent a new label).
3. If the participant **disputes** a resolution, mark it for review and exclude pending resolution.
4. Maintain a complete **audit trail**: every resolution records its method, evidence, adjudicator, and timestamp.

18 Scoring Methodology

All systems are scored with **proper scoring rules**, which are minimized in expectation only by reporting one's true probabilities — removing any incentive to game the metric through over- or under-confidence.

Primary scoring rule — logarithmic score. For a forecast distribution `p` and observed outcome `o`, the log score is `log p(o)`. Higher (less negative) is better. The log score is strictly proper and underlies the bit-denominated skill metric below.

Secondary scoring rule — Brier score. The multi-class Brier score, `sum_k (p_k - y_k)^2` where `y` is the one-hot outcome, is reported alongside the log score for robustness and to support calibration decomposition.

Personal Skill (bits). For a system `S` and reference `R`:

```
PersonalSkill(S vs R) = ( mean[ log p_S(o) ] - mean[ log p_R(o) ] ) / ln(2)
```

i.e., the mean per-forecast log-score advantage of `S` over `R`, converted to **bits** (using log base 2 equivalently). Positive values mean `S` forecasts the participant's future better than the reference.

Interpretation. Positive Personal Skill **vs R2** means the system forecasts the participant's future better than that participant's own routine baseline — the result that matters most. Positive skill vs R1 only (but not R2) means the system has learned population structure, not the individual.

Reporting. For each system we report:

- Overall Personal Skill (vs R1 and vs R2).
- Per-task skill (T1–T6).
- Per-participant skill.
- Per-evidence-tier skill (L0, L1, L2).
- Confidence intervals (day-blocked bootstrap; §22).
- Calibration status (ECE, reliability diagram, Brier decomposition).
- Identity-permutation result.

19 Baselines

The credibility of every positive result depends entirely on the strength of the baselines. **A weak baseline manufactures fake skill.** Both baselines are therefore specified rigorously and must be implemented and validated before any evidence-based system is scored.

19.1 R1 — Population prior

Forecasts derived from **aggregate outcome frequencies across participants** (and, where defensible, from sensible population defaults), with **no participant-specific evidence**. R1 answers: "How well can you forecast this person knowing only how people in general behave?"

19.2 R2 — Personal routine baseline

Forecasts derived from **the participant's own prior behavior only**, using **strict walk-forward** estimation — at cutoff `t`, R2 may use only outcomes resolved before `t`. R2 answers: "How well can you forecast this person knowing only their own past routine?"

> **Critical requirement.** R2 must be **strong**. If L2 beats only a weak, naive R2, the result is meaningless. R2 is the bar that genuine person-specific modeling must clear.

Reference R2 implementation.

- **Recency-frequency model:** per-task empirical distribution over the answer space from the participant's own history.
- **Exponential decay:** recent outcomes weighted more heavily than older ones (tunable half-life, fixed before analysis).
- **Weekday / time-of-day adjustment:** where a task plausibly depends on day-of-week or time (e.g., attention allocation differs on weekends).
- **Task-specific routine estimation:** each task family gets its own routine model rather than a shared one.

R2 hyperparameters (decay half-life, smoothing) are fixed before analysis and recorded in the pre-registration to prevent baseline-weakening by tuning.

20 Calibration

A system must not be allowed to win through overconfident lucky guesses. Calibration analysis verifies that stated probabilities correspond to observed frequencies.

Measures.

- **Expected Calibration Error (ECE):** weighted average gap between confidence and accuracy across bins.
- **Reliability diagrams:** plotted per system and per task.
- **Confidence bins:** fixed bin edges, set before analysis.
- **Brier score and its calibration/refinement decomposition.**

Provisional thresholds.

ECE	Status
<= 0.10	Pass
0.10 – 0.20	Warning
> 0.20	Fail

These thresholds may be revised **after the pilot** in light of observed bin counts, but they must be **fixed before analysis** of any given round. A system that achieves high nominal skill but fails calibration is reported as failing, not as succeeding.

21 Identity-Permutation Test

The permutation test is the study's strongest guard against the illusion of person-specific understanding.

Purpose. Determine whether measured skill reflects modeling of *this* participant, or merely exploitation of population-level structure shared across participants.

Method.

1. Take forecasts made for participant `i`.
2. Score them against the outcomes of participant `j` (j not equal to i), restricted to task/time-compatible pairs.
3. Repeat across many `(i, j)` permutations to build a null distribution of "permuted skill."
4. Compare permuted skill to true (unpermuted) skill.

Expected behavior. For a genuinely person-specific system, **permuted skill should collapse toward zero** — forecasts tuned to person `i` should not predict person `j`. If permuted skill remains high, the system is using population structure, not person-specific evidence, and any "personal" claim is unsupported.

Pass criterion (fixed before analysis). Unpermuted Personal Skill must be **positive and substantially greater** than the permuted-skill distribution — concretely, true skill should exceed the 95th percentile of the permuted-skill null, and mean permuted skill should be statistically indistinguishable from zero. Exact thresholds are recorded in the pre-registration.

22 Statistical Analysis Plan

Primary endpoint. Personal Skill **vs R2** on **T4 (attention allocation)**.

Secondary endpoint. Personal Skill **vs R2** on **T6 (commitment follow-through)**.

Additional pre-specified analyses.

- Personal Skill vs R1 (all tasks).
- Evidence-tier deltas: L2 − L1, L1 − L0, L2 − L0.
- Per-task performance (T1–T6).
- Per-participant heterogeneity.
- Calibration (ECE, reliability, Brier decomposition).
- Identity-permutation collapse.

Inference methods.

- **Paired comparisons by forecast instance:** systems are compared on the *same* sealed forecasts/outcomes, so all comparisons are paired at the forecast level.
- **Day-blocked bootstrap:** confidence intervals are computed by resampling whole days (not individual forecasts) to respect within-day correlation and avoid overstating precision.
- **Participant-level sensitivity analysis:** the primary analysis is repeated with each participant removed in turn, to confirm no single participant drives the result.
- **No population generalization claims.**

Power and interpretation. Statistical power in this pilot derives from the number of **resolved forecasts** (hundreds per participant across tasks and days), not from the 5-person sample. Even so, the study is **exploratory**: results indicate whether the instrument works and whether effects are plausibly present, not their magnitude in any broader population. All p-values/intervals are interpreted as instrument-validation signals, not confirmatory population inference.

23 Software and Infrastructure

Software.

- pwm-bench scoring harness (schemas + proper scoring rules + permutation).
- Python, with pandas / numpy / scipy.
- pytest for testing critical infrastructure (scoring, cutoff enforcement, sealing).
- WhisperX / faster-whisper (transcription).
- pyannote.audio (diarization).
- Local storage / NAS.
- SQLite or Postgres (evidence + outcomes).
- Optional vector database (topic/entity retrieval).
- Optional LLMs (evidence extraction and forecasting), kept replaceable.

Folder structure (per participant).

```
participant_id/
  raw_audio/          # never leaves secure storage; never published
  transcripts/        # sensitive; private
  derived_evidence/   # JSONL evidence records
  forecasts/          # sealed, hashed forecast JSONL
  outcomes/           # resolved outcome JSONL
  excluded_segments/  # flagged-for-deletion / excluded
  audit_logs/         # hashes, cutoff checks, adjudication trail
```

Security.

- Encrypted storage at rest.
- Access limited to the researcher and the participant.
- **No raw data in version control.** No raw audio in any public repository.
- No participant names in filenames; **pseudonymous IDs only**.
- Clean separation of storage, compute, and client responsibilities.

Testing-first commitments. Before live data collection, the scoring harness, cutoff-enforcement check, forecast-sealing/hash verification, and permutation routine are unit-tested against synthetic data with known answers. This catches silent scoring errors and leakage before they can corrupt real results.

24 Privacy, Ethics, and Consent

This is the most consequential section of the protocol. Audio recording of daily life is high-sensitivity data, and the study is designed around minimizing and containing that risk.

Risks.

- Capture of private conversations.
- Capture of third-party voices without their consent.
- Sensitive personal information in transcripts.
- Workplace or legal recording constraints.
- Psychological discomfort from being recorded.
- Re-identification risk from derived data.
- Misuse of predictive systems built on personal data.

Mitigations.

- Informed consent before enrollment (Appendix A).
- Participant pause/delete rights at all times.
- Daily participant review with segment-level exclusion.
- Exclusion of sensitive segments before processing where flagged.
- Local-first processing wherever practical.
- Encryption at rest; access strictly limited.
- Pseudonymous IDs; no names in files.
- Aggregate-only reporting; no raw data release; no public transcripts.
- No intervention during the scoring window.
- No clinical, medical, diagnostic, or therapeutic claims.
- **IRB review recommended before any larger study.**

Third parties. Participants must **not** record in settings where others have a reasonable expectation of privacy unless consent is obtained. Incidental third-party capture should be minimized and may be flagged for deletion.

Legal note. Audio recording laws vary by jurisdiction, including **one-party vs. two-party (all-party) consent** rules. The study must comply with the applicable laws in each participant's location, and each participant will be briefed on the rules that apply to them. This protocol does not constitute legal advice; where recording legality is uncertain, the participant is excluded or restricted to clearly lawful contexts.

Status of review. This pilot is small and exploratory. IRB or equivalent ethics review is **recommended** and is treated as a prerequisite before any expansion beyond this pilot or any involvement of clinical populations.

25 Participant Instructions

Operational, step-by-step. A standalone version appears in Appendix A-handout.

Starting a recording. At the start of your day, open your recorder/app and start recording. Confirm it is capturing audio. Keep the device on you or in your working area.

When to pause. Pause whenever you enter a private setting, a setting where recording is not allowed, or a conversation with people who have not agreed to be recorded. When in doubt, pause.

What not to record. Do not record in prohibited locations (e.g., certain workplaces, secured or clinical facilities), and do not record private conversations of others without consent.

Transferring files. At the end of the day, transfer your audio to the secure folder/NAS following the setup instructions. Do not email or upload raw audio anywhere else.

Daily review (5–10 min). Open the daily review form (Appendix B). Confirm major events, confirm which commitments were completed, confirm where your attention went, and flag any segment you want deleted or excluded.

Requesting deletion. You can request deletion of any segment or of all your data at any time, with no explanation required.

If sensitive information is captured. Flag the segment in the daily review (or tell the researcher) and it will be excluded/deleted. You do not need to listen back through it yourself.

Who to contact. Contact the researcher (Yuri Sylvester) using the contact details provided at enrollment for any question, concern, or withdrawal request.

Time burden. Recording runs in the background. Expect about **5–10 minutes/day** of active work for the review, plus a one-time setup of ~30–45 minutes.

26 Researcher Operating Checklist

Daily.

- ☐ Confirm audio received for each active participant.
- ☐ Run transcription.
- ☐ Run diarization.
- ☐ Extract evidence (versioned prompts).
- ☐ **Verify evidence cutoff** (no post-cutoff timestamps in forecast inputs).
- ☐ Generate forecasts (all systems, identical task defs).
- ☐ **Hash and seal** forecasts; store before any outcome.
- ☐ Write audit-log entry (hashes, cutoff check result).
- ☐ Resolve prior-day outcomes.
- ☐ Update exclusion log.
- ☐ Note participant burden / issues.

Weekly.

- ☐ Check for missing or low-coverage data.
- ☐ Inspect a sample of transcripts for quality.
- ☐ **Verify no future leakage** (audit hashes + timestamp scan).
- ☐ Review participant concerns / withdrawal signals.
- ☐ Back up encrypted files.
- ☐ Re-run scoring-harness self-tests.

27 Risks and Mitigations

Risk	Severity	Mitigation
Participant drops out	High	Over-recruit awareness; low daily burden; flexible pause; treat dropout data per analysis plan
Recording burden too high	High	Background capture; 4 h/day minimum; short review; check burden weekly
Poor transcription quality	Medium	Use strong local ASR; sample-inspect weekly; flag low-quality days
Diarization errors	Medium	participant/other/unknown labels; manual spot-checks; treat speaker as soft signal
Ambiguous outcomes	Medium	Strict exclusion rules; frozen answer spaces; audit trail
Model future-leakage	High	Hard cutoff; timestamp scan; sealing + hash verification; automatic violation flag
Overfitting to routine	Medium	Strong walk-forward R2; permutation test; held-out interpretation
Privacy incident	High	Local-first; encryption; pseudonyms; no raw data release; delete-on-request
Third-party consent issue	High	No recording where others expect privacy; minimize incidental capture; exclusion
No system beats R2	Medium	Pre-declared as an acceptable, informative outcome; triggers protocol revision, not data dredging

28 Abandonment / Failure Criteria

The pilot **fails to demonstrate PWM** (i.e., does not provide evidence for person-specific forecasting skill) if any of the following hold:

- No system beats **R1**.
- No system beats **R2**.
- Skill **does not collapse** under identity permutation.
- Calibration **fails badly** (systematic ECE in the fail band).
- Outcome **ambiguity is too high** (excessive unresolved/excluded rate undermines scoring).
- Participant **adherence is too low** (insufficient resolved forecasts).
- **Privacy concerns** make the protocol untenable in practice.

A failed pilot is still informative. Each failure mode maps to a concrete repair: a too-weak signal implies revising evidence extraction or task selection; failure to collapse under permutation implies the scoring or baselines are leaking population structure; high ambiguity implies tightening answer spaces and resolution rules; low adherence implies reducing burden. The pilot's job is to find these problems cheaply, before a larger study commits resources and additional participants' privacy.

29 Expected Outputs

Private artifacts (never published).

- Participant-level resolved-forecast datasets.
- Raw audio, transcripts, and derived evidence.

Aggregate / shareable artifacts.

- Aggregate scores (Personal Skill vs R1 and R2, per task and tier).
- Calibration plots (reliability diagrams, ECE).
- Evidence-tier comparison (L0 vs L1 vs L2).
- Identity-permutation results.
- Pilot report (feasibility, scoring validation, lessons).
- Revised protocol for a potential larger round.
- Possible preprint.

Public outputs (strictly limited).

- **No** raw audio.
- **No** private transcripts.
- Aggregate metrics only.
- Synthetic examples (no real participant content).
- Scoring code.
- This protocol.

30 Appendices

30.1 Appendix A — Sample Consent Language

Study: PWM-Pilot-Audio — A Prospective Evaluation of Person-Specific Forecasting from Longitudinal Audio Evidence.

Purpose. You are invited to take part in a research study about whether computer systems can forecast a person's near-future everyday choices (such as which project gets the most attention, or whether a stated commitment is followed through) using recordings of everyday audio, and whether such recordings improve predictions over simple baselines.

What you will do. For up to 30 days, you will record portions of your daily audio (target 8–12 hours/day, minimum 4 hours/day), transfer the files to a secure encrypted location, and complete a short daily review (about 5–10 minutes) confirming what happened and flagging anything to delete.

Data and privacy. Your raw audio will be stored encrypted, accessible only to you and the researcher, and will **never be shared or published**. From the audio, text transcripts and structured notes are created; only **aggregate, anonymized** results are reported. You are identified only by a code (e.g., P001). You may pause recording at any time and request deletion of any segment or all of your data.

Third parties. You agree not to record where other people reasonably expect privacy without their consent, and to minimize incidental recording of others.

Risks. Audio is sensitive and may capture private or third-party content. Recording laws vary by location; you will be briefed on the rules that apply to you. There may be mild discomfort associated with being recorded.

Not medical. This is research, not medical care, therapy, diagnosis, or coaching. No health assessment is made.

Voluntary. Participation is entirely voluntary. You may withdraw at any time, for any reason, without penalty, and request deletion of your data.

Consent. I have read and understood the above and agree to take part.

Name: _____ Signature: _____ Date: _____

30.2 Appendix A-handout — One-Page Participant Summary

(Provided as a separate standalone file: `PWM-Pilot-Audio_Participant_Summary.md` / `.pdf`.)

30.3 Appendix B — Daily Participant Review Form

```
Date: __________   Participant ID: P0__

1. Audio coverage today (hours, approx): ____
2. Major events today (brief):
   - ____________________________
3. Commitments stated earlier - outcome within window?
```

```
   [commitment]  -> completed / partial / not / superseded / unresolved
4. Where did most of your attention go today?
   ( ) PWM-Bench ( ) hardware ( ) family ( ) work/business
   ( ) health/admin ( ) other: ______
5. Any routine deviation today?
   ( ) none ( ) work ( ) family ( ) health ( ) travel ( ) social ( ) other
6. Segments to DELETE or EXCLUDE (time ranges + reason):
   - ____________________________
7. Anything sensitive captured that should be removed? (Y/N): __
8. Any concerns / device issues: ____________________________
```

30.4 Appendix C — Sample Forecast JSON

```json
{
  "forecast_id": "P001-T4-2026-06-06",
  "task": "T4",
  "participant_id": "P001",
  "system": "E-L2-audio",
  "cutoff": "2026-06-06T20:00:00-04:00",
  "sealed_at": "2026-06-06T20:07:42-04:00",
  "hash_sha256": "9f2c... (forecast payload hash)",
  "distribution": {
    "PWM-Bench": 0.46, "hardware": 0.08, "family": 0.12,
    "work/business": 0.22, "health/admin": 0.04, "other": 0.08
  },
  "cited_evidence": ["P001-2026-06-06-0042", "P001-2026-06-06-0051"],
  "prompt_version": "forecast-T4-v0.1"
}
```

30.5 Appendix D — Sample Outcome JSON

```json
{
  "forecast_id": "P001-T4-2026-06-06",
  "participant_id": "P001",
  "task_id": "T4",
  "resolution_time": "2026-06-08T09:15:00-04:00",
  "observed_answer": "PWM-Bench",
  "resolution_method": "transcript_topic_share+daily_review",
  "evidence_ids": ["P001-2026-06-07-0007", "P001-2026-06-07-0019"],
  "adjudicator": "researcher",
  "confidence": 0.9,
  "excluded": false,
  "exclusion_reason": null
}
```

30.6 Appendix E — Sample Evidence JSON

(See §14 for the full annotated record; one instance reproduced here.)

```
{
  "evidence_id": "P001-2026-06-06-0042",
  "participant_id": "P001",
  "timestamp_start": "2026-06-06T14:03:11-04:00",
  "timestamp_end": "2026-06-06T14:03:39-04:00",
  "source_type": "audio_transcript",
  "evidence_tier": "L2",
  "modality": "audio",
  "transcript_excerpt": "I need to send the grant draft to Dana before Monday.",
  "speaker": "participant",
  "evidence_type": "commitment",
  "entities": ["Dana", "grant draft"],
  "topics": ["work/business", "PWM-Bench"],
  "candidate_goal": "submit grant draft",
  "commitment": "send grant draft to Dana",
  "due_time": "2026-06-08T23:59:00-04:00",
  "attention_target": "work/business",
  "confidence": 0.82,
  "sensitivity_level": "normal",
  "excluded": false,
  "exclusion_reason": null,
  "model_used": "whisperx-large-v3 + extractor-v0.1",
  "prompt_version": "extract-commitment-v0.1"
}
```


30.7 Appendix F — Forecast Prompts

(Full text reproduced in §16: Prompts A–H. Each is versioned; the active version string is recorded in every forecast/evidence record.)

30.8 Appendix G — Evidence Extraction Prompts

(See §16 Prompt F and Prompt G; the extraction prompt enforces: factual-only, no clinical inference, no fabrication, schema-conformant output, confidence on uncertain items.)

30.9 Appendix H — Data Dictionary

Field	Type	Domain / Notes
participant_id	string	Pseudonymous (P0NN)
evidence_id / forecast_id	string	Unique, date-stamped
timestamp_* / cutoff / sealed_at	ISO-8601	With timezone offset
evidence_tier	enum	R1, R2, L0, L1, L2
modality	enum	audio, text, metadata

Field	Type	Domain / Notes
evidence_type	enum	commitment, decision, future_intention, concern, repeated_topic, attention_signal, routine_signal, social_interaction, event_update, uncertainty_signal
speaker	enum	participant, other, unknown
sensitivity_level	enum	normal, sensitive, restricted
task_id	enum	T1–T6
distribution	object	Probabilities over fixed answer space; sum=1
observed_answer	enum	One label from the task's answer space
resolution_method	string	Objective evidence and/or participant confirmation
adjudicator	enum	researcher, participant, participant-dispute
confidence	float	[0,1]
excluded	bool	—
exclusion_reason	string|null	Required when excluded=true
hash_sha256	string	Forecast payload hash (sealing)
prompt_version	string	Versioned prompt id

30.10 Appendix I — Exclusion Log Template

```
exclusion_id | participant_id | item_type(evidence/forecast/outcome)
| item_id | logged_at | reason | logged_before_scoring(Y/N) | adjudicator
```

30.11 Appendix J — Pre-Registration Checklist

- ☐ RQs, hypotheses, primary/secondary endpoints fixed.
- ☐ Task answer spaces frozen (per-participant T4 topic set recorded).
- ☐ R1 and R2 specifications + R2 hyperparameters fixed.
- ☐ Scoring rules, Personal-Skill definition, and bootstrap method fixed.
- ☐ Calibration measures and thresholds fixed.
- ☐ Permutation method and pass criterion fixed.
- ☐ Exclusion rules and "no post-hoc answer-space change" rule fixed.
- ☐ Cutoff time, sealing/hashing procedure fixed.
- ☐ Prompt versions frozen and archived.
- ☐ Privacy/consent procedures and legal review complete.
- ☐ Scoring harness self-tests passing on synthetic data.

30.12 Appendix K — Professor / Advisor Review Checklist

- ☐ Are baselines (especially R2) strong enough that beating them is meaningful?
- ☐ Is future leakage impossible by construction (cutoff + sealing + hash)?
- ☐ Are scoring rules proper, and is the skill metric correctly defined?
- ☐ Does the permutation test adequately establish person-specificity?
- ☐ Are answer spaces resolvable, and are exclusion rules pre-committed?
- ☐ Is the pilot honestly framed as instrument validation, not population inference?
- ☐ Are privacy, consent, third-party, and legal issues adequately handled?
- ☐ Are claims appropriately bounded (no understanding/clinical/medical overreach)?

31 Pre-Pilot Readiness: What Remains Before Running

(Summary of outstanding items; expands Appendix J.)

1. **Ethics/legal:** confirm recording-law status for each participant's jurisdiction; obtain IRB or equivalent review for any step beyond this pilot.
2. **Baselines:** implement and validate R1 and R2; freeze R2 hyperparameters.
3. **Harness:** implement and unit-test scoring, cutoff enforcement, sealing/hash verification, and permutation on synthetic data.
4. **Per-participant setup:** fix each participant's T4 topic set and answer spaces at enrollment.
5. **Pipeline dry-run:** end-to-end test (capture → transcript → evidence → forecast → seal → resolve → score) on one volunteer-day of synthetic or self-collected data.
6. **Prompts:** freeze and archive prompt versions.
7. **Pre-registration:** complete Appendix J and register before data collection.
8. **Consent:** finalize consent form with any institutional language; brief participants.

End of protocol. This is a pre-pilot protocol draft. It contains no empirical results and makes no medical, clinical, or diagnostic claims.