TARGETSPACE

Why TargetSpace

The engineering case for measuring what personal AI actually claims.

EDITORIAL · ~12 MINUTE READ · JULY 2026 · TARGETSPACE.ORG/WHY

> **Version 1.0 — pre-pilot protocol proposal.** Synthetic harness only; no human-subject results. This essay introduces the project for researchers, engineers, and builders, and assumes no familiarity with the paper. The paper — *TargetSpace: Benchmarking Personal Intelligence by Target-Specific Forecasting Under Partial Observation* — is the canonical reference wherever the two differ in precision.

1 A claim every product makes and no instrument checks

A growing class of products records, remembers, and reasons about one person over long stretches of time. Meeting recorders summarize your week. Wearable pendants transcribe your conversations. Glasses see what you see. Assistants accumulate months of email, calendar, and chat, and advertise the result as memory, context, or personal understanding. Whatever the form factor, the underlying promise is the same: *the longer this system observes you, the better it knows you.*

It is a testable promise. It is almost never tested.

What gets measured instead is everything around it: transcription accuracy, retrieval precision, summary quality, satisfaction, fluency. All real, all useful — and all able to be excellent while the promise itself is false, because none of them asks a question whose answer is not already in the record.

TargetSpace exists to ask that question. It is a benchmark protocol built around a single measurable capability: given longitudinal observations of one consenting person up to a sealed moment in time, can a system forecast what that specific person does next — better than population statistics, better than the person's own routine, with honest probabilities, in a way that falls apart when scored against the wrong person?

This page explains why that question, with those controls, is worth standardizing. The paper specifies the machinery; this is the reasoning that made the machinery necessary.

2 Where existing evaluations stop

Consider what today's evaluations ask. A transcript benchmark asks: *what was said?* A memory benchmark asks: *what happened?* A personalization benchmark asks: *what does this user prefer, given what they told us?* Each scores a system against evidence that already exists. Call the shared failure mode retrospective agreement: the answer key is the past, and the past is exactly what a large model with a large context window is best at re-serving.

Now consider what the promise actually implies. If a system has really built a model of you from months of observation, it should know things the record does not literally contain. Not just that you have a gym block on Thursdays — your calendar knows that — but that this Thursday's session will slip, because a deadline moved on Tuesday and your last two evenings went to the project. The first fact is retrieval. The second is a forecast, and it has a property retrieval never has: it can be scored against reality before reality happens, and it can be wrong.

> *Recall is graded against the past, which the system has already seen. A forecast is graded against the future, which nobody has.*

The difficulty is that a convincing forecast can be a fake in three ways, and a useful benchmark has to rule out all three.

Base rates. Most scheduled meetings happen. Most emails eventually get a reply. A system that predicts the population's average behaviour looks accurate while knowing nothing about you. The population-prior baseline, R1, removes this: skill counts only above what base rates already deliver.

Routine. You check email at nine. You defer the same recurring review most months. A model that memorizes your habits looks personalized while capturing only repetition — the subtle case, because most of anyone's life *is* routine. So TargetSpace fits an explicit model of your own routine, R2, from your own history, and demands skill beyond it. Beating R2 is the line between replaying the target's routine and demonstrating predictive skill specific to the target.

Wrong-target generality. Generic pattern-matching dressed up as intimacy. If a system's forecasts are equally good when scored against a different person's outcomes, they were never about you. Every system is also scored against wrong-target outcomes on matched tasks — the permutation gate — and genuine person-specific skill must degrade by a pre-registered margin when the pairing is broken.

Calibration holds the rest together. A forecast is a probability, and downstream decisions consume probabilities, so a system that says 80% must be right about four times in five. Overconfident luck fails the gate even when the top guesses are correct.

None of these instruments is new on its own. Proper scoring rules have graded weather forecasters since the 1950s; calibration analysis, prospective sealing, and ablation studies are standard practice; ubiquitous-computing research was fitting individual-routine models from passive data two decades ago. What has not existed is the conjunction, standardized: sealed, proper-scored forecasts about a tracked person, under an own-routine baseline and a wrong-target gate, with the evidence disclosed and attributed. That conjunction is the benchmark.

3 Longitudinal observation is a different problem

Static benchmarks get to ignore time. A labeled image is as good today as yesterday; shuffle the dataset and nothing breaks. Evidence about a person is nothing like this, and the differences are exactly where naive evaluation goes wrong.

Time cannot be shuffled

Chronology is load-bearing. Whether a commitment slips depends on what happened Tuesday, so evaluation must be walk-forward — the model sees only what was observable before the sealed moment, indices rebuilt per slice, no random cross-validation. People also drift: the routine baseline that was true in March misleads in June, so baselines are refit as time advances, and adaptation is part of what gets rewarded.

Missingness is evidence

Missingness carries information. When a wearable sits on its charger, the gap in the record is not random — charging hours correlate with exactly the schedule breaks a forecaster most needs to see. When a device is removed at dinner, that is consent working as intended, and the record must say so rather than silently impute. TargetSpace treats coverage logs and missingness mani-

fests as first-class artifacts: *no observation* and *no event* must remain distinguishable, or calibration analysis quietly rots.

Observation changes behavior

Observation is also reactive. A device the wearer performs for measures the performance, not the person. And the observed record is never the person: speech, location, calendar entries, and self-reports are partial projections of an underlying state through particular channels. A calendar entry is not a commitment; a missed reply is not avoidance. The record is not the target — which is precisely why the test must be whether a system can forecast the record's *next* entries, not summarize its existing ones.

Put together, this means personal-AI evaluation is incomplete when model scores are reported without characterizing the evidence behind them. Two systems with identical architectures and different capture coverage are different experiments. TargetSpace therefore treats observation quality — what was sensed, when, how completely, under what consent — as a first-class experimental variable, disclosed in machine-readable manifests and ablated like any other variable. The limiting factor for this field is increasingly not only model capability but the quality, continuity, and governance of the evidence a model is given. That deserves measurement, not assumption.

4 From memory to measurement

A fair objection: isn't this just asking whether the model *understands* the user? Deliberately, no — and the distinction is the project's spine rather than a lawyer's footnote.

Understanding, in the rich sense, is not observable. No benchmark measures it, including this one. What is observable is whether a system's stated probabilities about a specific person's future observable behaviour are better than base rates, better than that person's routine, honest about their own uncertainty, and specific to that person. TargetSpace names that bundle *target-specific predictive capability* and measures exactly it. Wherever the project uses the word understanding, it is operational shorthand for that bundle and nothing more — the same discipline by which prediction and explanation are kept distinct in statistics.

Narrowing the claim is what makes the claim testable, and testable in both directions. A well-run TargetSpace evaluation can return answers a marketing benchmark cannot afford: no lift over base rates; lift over base rates but nothing beyond routine; skill that fails calibration; skill that survives against the wrong person and was therefore never personal; a cheaper sensor matching

an expensive one. Every one of these nulls is an informative result about a named configuration, and the protocol is built so they are reportable rather than buried. A field that cannot publish its nulls cannot converge.

This is also why the benchmark is an instrument, not a philosophy. ImageNet did not define vision; it standardized a measurement that let a field converge on what worked. SWE-bench did the same for software agents. TargetSpace claims the same functional role for personal AI — the shared task that turns a diffuse promise into a number with error bars — and no comparable scale or importance. The larger inquiry it serves, which the paper calls human observation science as a *proposed* research program, is the study of which observations, schedules, and governance constraints are sufficient for reliable person-specific prediction. The benchmark is how that program gets its data.

5 More sensors are not automatically better

A benchmark that rewarded raw predictive skill alone would carry an ugly incentive: capture everything, always, from every angle, because somewhere in the pile there might be bits. Personal AI does not need an instrument that grades surveillance as diligence.

TargetSpace is built so that invasiveness has to pay rent. Two mechanisms do the work.

Evidence efficiency asks, of every added sensor, modality, or sampling schedule: how much gated skill did it add, per unit of what it cost? Cost is measured in the units engineers actually budget — joules, worn hours, captured bytes, required manual interactions — and only skill that survives every gate counts in the numerator, so a stream cannot buy efficiency with noise. Privacy is deliberately never collapsed into the denominator as a single number; it stays a reported vector — raw audio hours, image counts, retention, bystander exposure — because a scalar privacy score is how privacy gets traded away.

Minimum sufficient observation is the design objective the ratios point at: the least burdensome governed evidence configuration that still preserves a pre-specified level of gated skill. It is task-dependent by construction. For forecasting whether a meeting slips, calendar metadata may be enough. For forecasting a shift in what someone is working toward, audio may earn its place. For some outcomes, no acceptable configuration may exist at all — and that finding is a result, not a failure. Where two configurations tie within a pre-registered margin, the less invasive one wins the comparison by rule.

> *The benchmark's question is never "what could we capture?" It is "what is the least we can observe and still know what we claim to know?"*

This inverts the field's default. Today, sensing decisions are made by product intuition and component pricing; justification, if any, comes later. Under measurement, an engineer can defend a microphone the way they defend a cache: with a curve. And a regulator, an IRB, or a user can ask for the same curve.

6 The engineering surface

Framed this way, personal AI stops being one product race and becomes several measurable engineering problems, each with its own community.

Hardware. No shipping device was designed as an observation instrument; today's wearables spend their budgets on interaction. The open design space — duty cycling against battery-shaped missingness, sensor placement against attribution quality, edge preprocessing against evidence loss, disclosure hardware that bystanders can see — currently runs on intuition. Bits of gated skill per joule, per worn hour, per captured byte turn it into an optimization with a scoreboard. The paper's ideal observation device is stated as a testable envelope, not a product spec: every property, from battery life that outlasts the day to the absence of a display, traces to a validity requirement you can check.

Software and protocol. The reference harness, schemas, and baselines are deliberately boring: deterministic synthetic data, machine-readable manifests for evidence, hardware, coverage, and energy, a scoring pipeline that refuses to headline any efficiency number whose gates did not pass. Sealing is a protocol, not a pinky-swear — forecast hashes must reach an external witness before outcomes exist, and runs without one are labelled self-attested. Boring is the point; reproducibility lives in the boring parts.

Governance. Federation is the protocol's specified path for personal data: raw recordings stay with the participant or the company that holds them, the harness travels to the data, and what crosses the boundary is sealed forecasts, resolved outcomes, and aggregate reports above a minimum cohort size. The trust model is explicit rather than implied — a self-run evaluation is a within-study diagnostic, and only organizer-verified runs compare across systems. Consent is ongoing, bystander exclusion is a hard constraint, and a valid forecast never grants permission to act

on the person. These are design commitments of a pre-pilot protocol, stated so they can be audited, not achievements to date.

Ecosystem. Meeting tools, passive audio, glasses, wearables, enterprise copilots, assistive systems — the categories look unrelated until you notice they are climbing the same ladder: recording, then structured memory, then longitudinal evidence, then — if it exists at all — person-specific predictive capability. A shared instrument at that rung gives every category the same starter experiment: pick a high-frequency observable outcome, run the same sealed tasks with your feature on and off, and report whether memory became foresight.

7 What TargetSpace actually contributes

Stripped to its parts, the contribution is four named things, offered at pre-pilot status with a synthetic reference harness and no human results — the paper is explicit about this boundary, and so is every page of this site.

	NAMED THING	WHAT IT IS
INSTRUMENT	**TargetSpace**	Sealed, proper-scored forecasts of one tracked person's future observable transitions, gated by R1, R2, calibration, and wrong-target permutation, with evidence disclosed and attributed.
MEASURED OBJECT	**Target-specific predictive capability**	What a passed evaluation certifies — and the operational content behind any use of the word “understanding.”
OPTIMIZATION METRIC	**Evidence efficiency**	A family of gated ratios — incremental skill per disclosed unit of observation cost — connecting sensing decisions to measured value.
DESIGN OBJECTIVE	**Minimum sufficient observation**	The least burdensome governed configuration that preserves a specified skill level. Less invasive wins ties, by rule.

Equally important is what is *not* claimed. The ingredients — proper scoring, calibration, sealing, ablation, routine models from passive data — are adopted from decades of prior work in forecasting, evaluation, and ubiquitous computing, and the paper cites its debts. No architecture is mandated: a personal world model is one hypothesis for passing the benchmark, not a requirement of it. And nothing here proves that passive observation improves forecasting at all — that is the flagship hypothesis the instrument exists to test, and it is allowed to come back false.

8 Where this could lead

What follows is direction, not result; none of it is established, and the paper's claims table governs anything this section seems to promise.

If the first studies validate — if passive longitudinal evidence produces calibrated, person-specific lift beyond routine for even a few task families — then observation architecture becomes a measurable discipline. Hardware gets evaluated by skill-per-joule curves instead of spec sheets. Product claims about memory become auditable with a two-arm experiment any team can run in weeks. Categories that today compete on demos acquire a common yardstick, and the phrase “our AI knows you” starts arriving with a number, a margin, and a disclosed evidence bill.

If the studies come back null — if nothing beats the routine baseline, or richer capture adds nothing metadata didn't already carry — that is not the project failing. That is the project working: a public, specific, reproducible answer to a question the industry currently answers with adjectives, and a redirect of engineering effort away from capture that cannot pay for itself.

Either way, the field gets something it does not have today: a way to be wrong in public about a claim that is currently unfalsifiable in private.

Personal AI has spent a decade promising that it understands people. TargetSpace is the narrower engineering step of asking that promise to make a forecast — sealed, scored, specific to you, and accountable to the future.

FURTHER READING

The paper (49 pp, Version 1.0) `targetspace.org/paper`

Run a starter experiment `targetspace.org/product-quickstart`

Observation science `targetspace.org/observation-science`

The protocol `targetspace.org/protocol`

Code, schemas, synthetic harness `github.com/yurisyl/targetspace-bench`