TargetSpace: Benchmarking Target-Specific Understanding by Prospective State-Transition Prediction

A benchmark protocol and validity stack for testing whether AI systems form calibrated, target-specific predictive models under partial longitudinal observation — with own-routine baselines, wrong-target controls, calibration, evidence ablation, and evidence-cost reporting

Yuri Andrade Sylvester
Independent Researcher
Carlsbad, California, United States
yurisyl@gmail.com
ORCID: 0009-0008-6513-6087

Public preprint — Version 1.2 (pre-pilot protocol proposal)

Abstract

AI systems increasingly claim to remember, personalize, act as agents, and model the people, teams, and organizations they serve. Existing benchmarks measure general knowledge, static tasks, conversational quality, tool use, code, or retrospective retrieval; none adequately tests whether a system has learned a useful model of a specific target over time. TargetSpace is a benchmark protocol for that question. Its anchor metric is prospective state-transition prediction: a system observes a consenting target — a person, team, organization, patient, project, device, or environment — up to a sealed time t and commits calibrated probabilistic predictions of the target's future state transitions before outcomes exist; outcomes resolve by deterministic rules and are scored with strictly proper rules. A validity stack determines whether the prediction is meaningful: skill must exceed a population prior (R1), the target's own routine (R2), and retrieval-and-summarization over the same evidence; remain calibrated; collapse under wrong-target substitution; concentrate on regime shifts rather than routine continuation; and attribute its lift to disclosed evidence streams. An observation-science layer prices that evidence, reporting predictive lift against volume, sensitivity, collection burden, energy, latency, consent burden, and privacy risk, with minimum sufficient observation as the design objective. Understanding is not reducible to prediction; TargetSpace operationalizes one measurable component of it — whether a system forms a calibrated, target-specific model with predictive force over time — and surrounds that observable with the controls that make it resistant to shallow correlation. Minimal, Standard, and Gold-Standard deployment levels define what early implementations must include and what the full apparatus adds. This is a pre-pilot protocol with a synthetic reference harness; no human-subject results are reported, and a high score certifies target-specific predictive force, not understanding in any richer sense, inner life, causation, or permission to act.

1 Introduction

AI systems increasingly claim to remember users, personalize over time, act as agents on someone's behalf, model teams and organizations, and adapt to the individuals they serve [25–32,49–55]. These claims share one structure: the system is said to have learned something about a *specific target* from longitudinal exposure. Yet the field's benchmarks measure almost everything except that. General knowledge, static tasks, conversational quality, tool use, code, and retrospective retrieval are well instrumented [1,2,5,6]; whether a system has formed a useful model of *this* user, *this* team, or *this* patient over *this* stretch of time is asserted in product language and measured almost nowhere. The claimed model is latent — it cannot be read off a transcript, a memory store, or a preference profile — so it can only be tested against what the target does next. Memory, retrieval, summarization, and preference replay each re-serve a record that already exists; the failure mode they share is retrospective agreement.[1]

We use *understanding* in a bounded, operational sense: understanding is the possession of a useful model of a target — a model that supports prediction, explanation, uncertainty estimation, counterfactual reasoning, and adaptation under changing conditions. TargetSpace focuses on the most directly measurable observable in that family: **prospective state-transition prediction**. Understanding is not reducible to prediction. But prospective state-transition prediction is one of the most rigorous observables for testing whether a model of a target has predictive force, and TargetSpace's contribution is to make that observable target-specific, longitudinal, calibrated, and resistant to shallow correlation (Sections 3 and 4). The benchmark tests a measurable proxy for target-specific understanding — predictive force over a specific target across time — and claims nothing richer.

The limiting factor is increasingly not only model capability but observation quality. A stronger model cannot indefinitely infer target-specific state from evidence that does not contain it, and richer evidence is not automatically useful: quality, continuity, attribution, coverage, and timing constrain what any model can learn about a particular person. Personal-AI evaluation is therefore incomplete when model performance is measured without characterizing the evidence available to the model. TargetSpace treats observation quality as a first-class experimental variable: every prediction is attributed to a disclosed evidence configuration, and the evidence-tier ablation measures what each observation stream contributes.

TargetSpace addresses this gap. It is a benchmark protocol for prospective state-transition prediction about a specific target: a system observes a target — a person, team, organization, patient, project, device, or environment — up to a sealed time, then commits calibrated probabilistic predictions of the target's future state transitions before outcomes exist. Predictions are scored with strictly proper rules against deterministically resolved outcomes, and skill counts only when it survives a **validity stack** (Section 7): it must beat generic priors (R1), beat the target's own routine (R2), beat retrieval-and-summarization over the same evidence, remain calibrated, collapse under wrong-target substitution, concentrate on regime shifts rather than routine continuation, and attribute its lift to disclosed evidence streams at disclosed cost. The claim is about standardization, not priority: ubiquitous-computing research has long fit individual-routine models from passive longitudinal data [66] and compared person-dependent with person-independent models and sensor subsets [65]; what the field lacks is a shared, prospective, sealed instrument that makes such claims comparable. ImageNet answered "what does progress in image recognition look like?" [1]; SWE-bench answered it for software engineering [2]. No shared artifact yet answers it for the central claim of memory, personalization, and agent systems — calibrated, target-specific prediction — and TargetSpace is proposed in that functional role, claiming no comparable scale or importance (Figure 1). The purpose is not abstract philosophy; it is a practical measurement layer for AI systems that claim target-specific understanding, and Section 2 states what builders improve to

[1] Self-report remains one auxiliary evidence channel, not the substrate (Section 5); the contrast implies no claim of access to consciousness or subjective states.

score higher — anywhere in the loop from observation to validation, with no single architecture prescribed.

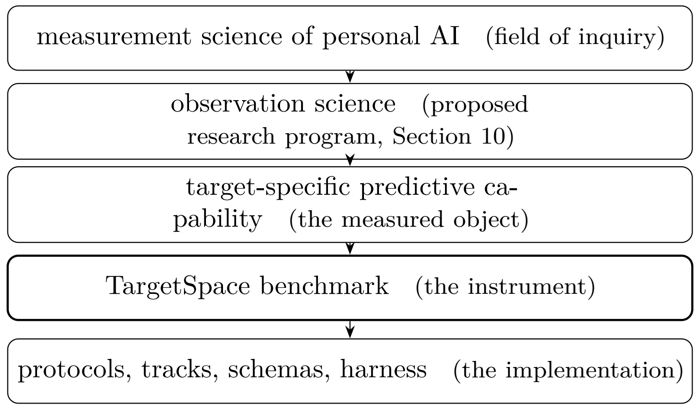

Figure 1: The conceptual hierarchy. The benchmark is the operational instrument inside a larger program; it is not itself the science, and the science is not reducible to it.

The benchmark rests on a simple distinction: the record is not the target. Personal AI systems observe traces — speech, location, messages, calendar entries, tasks, physiological signals, and self-reports — but each trace is a partial observation of an underlying target-state through a particular channel, not the target-state itself. A calendar entry is not a commitment, a missed reply is not avoidance, and a late-night search is not a goal; each becomes scorable only when mapped to a future observable resolution rule. A system that retrieves or summarizes these traces can reproduce the record without modeling the state that generated it. TargetSpace therefore tests whether a system can infer enough of the latent target-state to predict future observable transitions under sealed, proper-scored conditions.

The paper proceeds as follows: the commercial and scientific need (Section 2); why prediction is the anchor observable, and when it fails as a proxy (Sections 3–4); the task, the validity stack, and deployment levels (Sections 6–7); the flagship personal track and evaluation protocol (Sections 8–9); observation science and evidence cost (Section 10); conceptual lineage and related work (Section 12); and the release roadmap (Section 9.8). Throughout, predictive and explanatory goals are kept distinct in the standard sense [74].

The difficulty a benchmark must solve is that convincing predictions need not be target-specific at all. A model can look skilled by predicting what usually happens (most emails get answered) and personalized by replaying a target's routine (this person checks email at 9am), while capturing nothing about the individual beyond base rates and habit. Target-specific skill is what remains after both are subtracted: a prediction that beats a strong own-routine baseline, stays calibrated, and loses its skill when scored against the wrong target. The target-specific question is whether the model knew this person would ignore a particular message because their attention had shifted this week. We use the term personal intelligence as a bounded, product-facing label for this capability in the personal track, operationalized entirely by the prediction task and carrying no claim of general intelligence or access to inner life (and no relation to the personal-intelligence construct in personality psychology [76]).

TargetSpace operationalizes this test. A system receives observations of a target up to a sealed time t and emits calibrated probabilistic predictions of the target's future target-state transitions before the outcomes exist; predictions are timestamped and hashed, outcomes resolve by deterministic rules, and scoring is strictly proper; with seals committed to an external witness before outcomes exist, the design is contamination-resistant by construction, and without one a run is self-attested (Section 6.3). The target-specificity stack makes target-specificity the measured quantity (Section 7): sealed proper-scored prediction with the R1 population prior, the R2 own-routine baseline, a calibration gate, a permutation specificity gate, and an evidence-tier ablation that measures what passive observation adds over digital exhaust and self-report rather than assuming passive capture is superior. This paper specifies the task, the controls, and a synthetic worked example, and reports no empirical results. Because the object of evaluation is the sealed prediction and its resolved outcome, not any public release of raw capture, TargetSpace can run in a federated mode: raw longitudinal data stays under the participant or data holder's control, and only sealed predictions, resolved outcomes, and aggregate metrics are exported (Section 13).

A concrete instance makes this precise. A note-taking app claims that remembering a user makes it more helpful. TargetSpace turns that claim into a sealed prediction: before the outcome exists, the app must predict whether the user will complete, defer, cancel, or replace a specific recurring commitment. The prediction earns credit only if it beats the population completion rate (R1), beats the user's own routine (R2), stays calibrated, and, rescored against another user's outcome, loses its skill. A memory feature that cannot clear these four bars is performing retrieval, not modeling the user. This is the paper's minimal runnable task (Section 8.1), and every personal-track task follows its shape.

The formulation applies to any target for which repeated observations and deterministically resolved outcomes make an own-routine baseline and a wrong-target control well-defined: a person, a team, an organization, a patient, a project, a device, or an environment. The personal track (TS-Personal, "personal intelligence" as a bounded product-facing label) is the flagship and the only track instantiated here, with no claim of cross-domain empirical validation (extension criteria are in Appendix L).

1.1 Contributions

This paper contributes a benchmark apparatus and the research program built around it. Most individual tools are adopted from prior work and cited; the audited novelty is their conjunction around a single measurable question — has this system learned this target? — not any ingredient in isolation (Sections 7, 12). **(1) A prospective, contamination-resistant benchmark** for target-specific state-transition prediction under partial observation: sealed, proper-scored predictions of a tracked target's future observable transitions — proposed here (Sections 6–7). **(2) The validity stack**, assembled from established tools around two anchoring layers — the admitted R2 own-routine baseline and the wrong-target control — with sealing, proper scoring, R1, retrieval-only comparison, calibration, regime-shift stratification, evidence ablation, and evidence-cost reporting, plus advanced modules for counterfactual probes, out-of-distribution degradation, explanation faithfulness, and optional internal-representation probes (Section 7). **(3) An architecture-by-evidence evaluation grid** attributing predictive lift to the observations that produced it (Section 7; Appendix L). **(4) Observation science as an evaluation layer**: evidence efficiency, the model–evidence frontier, and minimum sufficient observation, pricing predictive lift against evidence volume, sensitivity, burden, energy, latency, consent, and privacy risk (Sections 10.3–10.4). **(5) A deployable protocol**: Minimal, Standard, and Gold-Standard levels, a concrete evaluation protocol card, the hypothesis ladder H0–H7, and a versioned release roadmap (Sections 7.2, L.2, 9.8). **(6) Implemented artifacts**: federated starter protocols, machine-readable schemas, cross-task transfer specification, and a synthetic reference harness (Appendices B, C; Section 8.4).

For reference, Table 1 states what this pre-pilot paper claims and does not claim.

Table 1: What this pre-pilot paper does and does not claim.

Item	Status
TargetSpace benchmark framework	Claimed (specified here)
The personal track (TS-Personal) / personal intelligence	Specified as flagship track; "personal intelligence" is a bounded, product-facing label for that track's target capability, operationalized only as the prediction task — not a claim of general intelligence
Self-report as auxiliary evidence, not the substrate	Framing claim (Section 5)
Synthetic demonstration harness	Implemented; sanity check only
Human pilot results	Not claimed (no pilot run)
Cross-domain empirical validation	Not claimed
Passive observation improves prediction	Hypothesis; to be tested by evidence ablation
Attention causes target formation	Not claimed; tested for incremental predictive value
Public raw first-person dataset	Staged, consent-governed release planned; not in initial release
Understanding, consciousness, or unbiased capture	Not claimed; a high score certifies calibrated predictive skill about observable future states only, and passive capture carries measurable, not absent, bias
Deployment legitimacy / permission to act on predictions	Not claimed

2 Why this benchmark now: the commercial and scientific need

Memory and personalization are becoming product claims across AI assistants, copilots, agents, healthcare AI, enterprise AI, education AI, wearables, and robotics. Without a benchmark, those claims are evaluated through demos, anecdotes, subjective preference, or retrospective examples. Companies need to know whether adding memory, sensors, context, or longitudinal data actually improves target-specific prediction; labs need a way to compare architectures, memory systems, context windows, retrieval methods, agent traces, multimodal observation streams, and privacy constraints on identical sealed instances; regulators, enterprise buyers, and users need evidence that systems improve through meaningful target modeling rather than through more surveillance or more stored context. TargetSpace is the evaluation layer for that market: it sells the ruler, not another assistant. Generic AI evaluations ask whether a model knows the world; TargetSpace asks whether a model has learned *this* target, and whether longitudinal evidence creates measurable target-specific lift.

What a higher score means for builders. A benchmark changes a field by defining what improvement means, and TargetSpace defines improvement as validated target-specific predictive lift. That definition does not prescribe an architecture or a single algorithmic path; it puts pressure on the whole loop — observation, representation, prediction, calibration, evidence minimization, validation. A system's score can rise because it collects more useful evidence, learns better temporal representations of target state, calibrates its uncertainty, identifies which evidence streams actually matter, reaches the same gated skill at lower evidence cost, or stays robust when the target leaves routine. The stack equally fixes what a score cannot be improved by: stored memory that never becomes prospective skill fails against R2 and the retrieval-only arm; more invasive capture that adds no gated lift loses under minimum sufficient observation; and scale that only sharpens population priors fails the wrong-target control. TargetSpace forces progress in target modeling, not merely model scaling, memory storage, or data capture: it rewards any system that can transform longitudinal observation into prospective, calibrated, target-specific prediction beyond routine and retrieval.

Today's personal-AI products look fragmented. Meeting recorders, AI wearables, camera-and-microphone glasses, memory assistants, digital companions, enterprise copilots, passive audio devices, and lifelogging systems ship as separate categories with separate metrics [49–55]. Their trajectories are not separate. Each starts from a bounded capture problem, accumulates longitudinal context, and advertises the same destination: a persistent model of a particular person that improves what the product does next. The market vocabulary of memory, context, and goals is advancing faster than any way to test it (the industry survey is Appendix M).

Figure 2 names the shared ladder. Recording externalizes single events. Structured memory persists them. Longitudinal evidence links them across weeks. Target-specific predictive capability turns them into calibrated predictions of what this person does next, and validated assistance is what that capability can epistemically support (deployment legitimacy is a separate question, Section 13). The ladder climbs toward a capability, not a representation: a personal world model is one architectural hypothesis for the internal state; target-specific prospective skill is the measured capability. TargetSpace requires no world model, no memory graph, no simulator, and no particular architecture — a reactive policy, a state-space model, an LLM, a JEPA-style latent predictor, or a probabilistic program may all participate, provided each produces sealed probabilistic predictions under the protocol. The claim is observational: it is evidence that the capability class is arriving, not validation of any method, not a claim that any vendor misuses data, and not a claim that any vendor has reached, or should reach, the top rung.

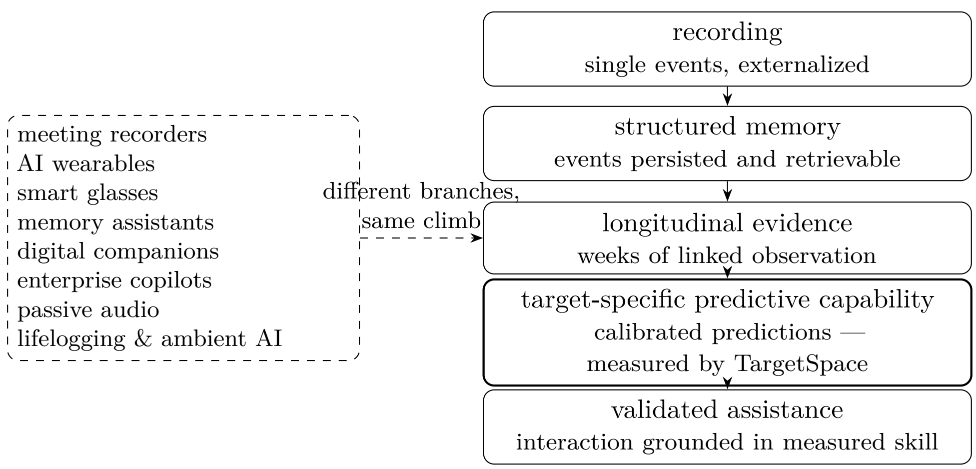

Figure 2: The convergence ladder. Today's product categories (left) start from different capture problems but climb the same ladder (right): recording becomes structured memory, memory becomes longitudinal evidence, and evidence either does or does not become target-specific predictive capability — the rung TargetSpace measures. The internal representation is unconstrained; the placement of categories is illustrative, not a ranking.

2.1 Three layers, one sequence

The program separates into three layers with different time scales and different owners.

Layer 1 is scientific instrumentation. Can a given evidence stream produce calibrated, target-specific predictive lift beyond the R1 population prior and the R2 own-routine baseline? Today's hardware — phones, recorders, wearables, glasses, and enterprise capture systems used as instruments

— is sufficient to pose the question. A device can be a poor all-day product and still be a good instrument; its role here is experimental (Section 10).

Layer 2 is observation architecture. If Layer 1 validates, the question becomes how to acquire faithful, low-burden, temporally useful evidence over months to years. Its objective is not interaction; it is evidence acquisition under constraints — modality, duty cycle, placement, timestamp quality, coverage, comfort, charging, buffering, exportability, missingness, privacy controls, reactivity, and energy. Section 10 states it as an open research problem, not a product roadmap.

Layer 3 is interactive personal AI. Assistants, displays, chat, recommendations, reminders, and agents are interaction surfaces. In the sequence this paper argues for, they are downstream applications built on validated predictive capability, not the place where that capability is assumed.

The sequence is observe, infer, forecast, validate, interact (Figure 3). Observation produces evidence; inference maintains a belief state over the latent target-state; forecasting commits to a distribution before the outcome exists; validation scores it against the sealed outcome; interaction then acts on a model that has earned its claims. Interaction quality is a legitimate product metric; it is not evidence of a target-specific model, which is why the benchmark scores the prediction and nothing else.

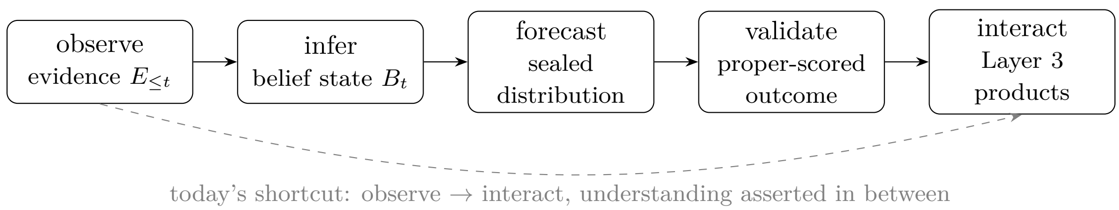

Figure 3: One sequence, two paths. The benchmark implements the validate step: the point where a claimed model of the person either survives contact with the future or does not.

2.2 What TargetSpace lets builders decide

The same apparatus that answers the scientific question answers a set of product and architecture questions, by turning each into a controlled comparison rather than a matter of conviction. A team runs the same sealed task set with one factor changed — a feature toggled, a sensor added, a model swapped, a window shortened — and reads the change in gated, target-specific skill and its disclosed cost. Table 2 lists representative decisions and the control that settles each; every entry reduces to *Skill over R2, under the gates, at a disclosed cost*, never interaction quality or offline accuracy. Because the comparisons run in the federated mode (Section 6.3), a team can answer them on its own users' data without exporting raw evidence.

3 Why prediction?

Several observables could anchor a benchmark of target-specific understanding, and none is sufficient alone. Table 3 compares the candidates. Prospective state-transition prediction is chosen as the anchor not because it exhausts understanding, but because it is prospective, falsifiable, naturally quantitative, comparable across systems and architectures — including black-box commercial systems — deployable in commercial settings, and directly tied to whether a model of a target has predictive force. The alternatives each fail as an *anchor* while remaining valuable as modules: intervention and control are often stronger evidence but are risky, ethically constrained, confounded

Table 2: What TargetSpace lets builders decide. Every answer is a change in gated, target-specific skill at a disclosed cost — never interaction quality; ties within the pre-registered equivalence margin rank the cheaper or less invasive configuration first (Section 10.4).

Decision	Control that answers it
Does long-term memory improve a model of the user?	run the sealed task with memory on/off; read Skill over R2. Memory that lifts R1 but not R2 is retrieval, not a personal model.
Does retrieval add predictive value?	toggle the retrieval feature across identical sealed instances; Skill over R2 plus calibration.
Does audio add value beyond calendar and text?	evidence-tier ablation L0/L1 → L2, reported as EE per disclosed cost (Section 10.2).
Does video add value beyond audio?	evidence-tier ablation L2 → L3–L4, same gating.
How much history is enough?	observation-window ablation (zero / short / longitudinal arms); the minimal window within the equivalence margin (Section 9.3).
Is a larger model better under the same evidence?	architecture comparison at a fixed evidence tier on the grid (Section 7).
Can inference stay on-device?	EE-energy with processing locality in the cost vector; edge versus cloud at equal gated skill.
Is the system calibrated enough to abstain?	the calibration gate plus abstention and coverage reporting (Section 9).
Which data can be deleted?	the minimal sufficient tier: streams outside the equivalence margin can be dropped without losing validated skill (Section 10.4).
Does personalization exceed routine replay?	Skill over R2 plus the routine-versus-transition split (Section 9.6).
Does the representation transfer across features?	cross-task personal transfer (Section 8.4, hypothesis H6).
Does better prediction improve downstream assistance?	the downstream-utility track (hypothesis H7), under a separate intervention and safety protocol (Section 13).

by execution, and hard to standardize; counterfactual reasoning is crucial but hard to verify without experiments or natural experiments; planning mixes understanding with optimization, goal selection, tool access, and execution quality; explanation is useful but vulnerable to post-hoc rationalization; compression and generalization are elegant but harder to communicate and operationalize; retrospective reconstruction and summarization are useful baselines but too easy to satisfy with retrieval; and internal-representation probing is valuable when internals are available, which for many commercial systems they are not. Prediction is the anchor; the validity stack (Section 7) determines whether the prediction is meaningful.

4 When prediction fails as a proxy

A prediction benchmark must not be naive about prediction. Six failure modes are standing objections. **Goodhart dynamics**: once a proxy becomes the optimization target, systems can improve the proxy without improving the capability it was meant to measure [87]. **Shortcut learning**: models exploit shallow correlations that happen to predict outcomes in-distribution [86]. **Causal blindness**: prediction from observation is associational, below intervention and

Table 3: Candidate observables for target-specific understanding. Prediction anchors the benchmark; the others enter as validity-stack modules where they can be verified.

Observable	**What it tests**	**Strength / limitation as anchor**	**Role in TargetSpace**
Prospective state-transition prediction	predictive force of the target model	falsifiable, quantitative, comparable, black-box-compatible / associational only	**anchor metric**
Intervention / control	causal grasp of target dynamics	strongest evidence / risky, ethically constrained, execution-confounded	optional module, where ethical
Counterfactual reasoning	model of unrealized alternatives	probes structure / hard to verify without experiments	advanced probe (Gold)
Planning	using the model to act	deployment-relevant / mixes modeling with optimization and tools	downstream track (H7)
Explanation	human-legible account of the model	interpretable / vulnerable to post-hoc rationalization	faithfulness check (Gold)
Compression / generalization	structure captured per bit	principled / hard to operationalize and communicate	implicit in transfer (H6)
Retrospective reconstruction	recall of the record	easy to run / satisfied by retrieval alone	baseline to beat
Internal-representation probing	what hidden states encode	mechanistic / requires white-box access	optional module

counterfactuals in the causal hierarchy [85]. **Predictive opacity**: accurate predictions need not yield human-understandable explanations [74]. **Distribution shift**: shortcut predictors collapse outside routine patterns. **Contamination and retrospective leakage**: without temporal sealing, apparent prediction is partly retrieval [3,4].

The TargetSpace validity stack is designed precisely to answer these risks. Sealing and walk-forward evaluation answer contamination. The R2 own-routine baseline and the retrieval-only comparison answer the cheapest shortcuts — habit replay and lookup. The wrong-target control answers correlations that are generic rather than target-specific. Regime-shift stratification and out-of-distribution degradation curves answer distribution shift: structural target modeling should degrade more gracefully than shortcut prediction when the target departs from routine. Explanation-faithfulness checks answer opacity where explanations are offered, and the counterfactual and intervention modules address — without claiming to close — the associational ceiling. Goodhart pressure is answered structurally rather than by exhortation: the stack is a conjunction, and optimizing any single layer while failing another downgrades the claim (Section 9.3, Table 7).

5 Evidence channels: self-report and passive observation

The apparatus of Section 7 is agnostic to how a target is observed; the personal track motivates a specific choice of evidence. Two channels are available — **first-person self-report** (prompts, journals, notes, saved memories, stated goals, surveys) and **third-person passive observation** (a longitudinal record of observable behaviour and context) — and neither grants access to "the real mind," which TargetSpace never claims: it scores only predictions of future observable states.

Neither channel is the target-state itself; each is an observation of it through an instrumented channel, so neither is a gold standard, and TargetSpace ranks channels by measured predictive lift rather than presuming one is ground truth. We reject any claim of observer superiority; the issue is **asymmetric observability**: self and observer see different parts of the state, their biases differ in kind, and their relative predictive value is an empirical question TargetSpace measures — by evidence tier, on identical sealed instances — rather than assumes.

Self-report is privileged but structurally biased. Self-report has genuinely privileged access to subjective experience, and we do not dismiss it; but it is not the target-state itself, only another observation channel, and as a substrate for prediction it is structurally biased in ways hard to remove: it is selective and post-hoc (autobiographical memory is reconstructed toward current goals [44], and retrospective accounts diverge from momentary experience [38]), introspection-limited [36], socially filtered [56,57], and declared intentions predict behaviour incompletely: intention–behaviour associations are moderate to large, yet a substantial fraction of intenders fail to act [45] — which makes stated goals a strong baseline and comparison arm C a stringent test, not a strawman.

This makes self-report an evidence modality and a baseline (human self-report is an architecture class in the grid, Appendix L), not the substrate or a gold standard.

Passive observation is biased differently, but more instrumentable. A longitudinal observer can detect regularities, inconsistencies between stated and enacted priorities, and slow changes hard to notice from inside experience; self and informants know different things, each more accurate for different trait classes [37]. Passive capture is not the target-state itself either, but an externally measured trace, and it is not unbiased: microphones, cameras, sampling windows, ASR and vision-model error, missing context, and reactivity to being recorded all bias the record.

The comparative claim is a hypothesis, not a structural fact: we hypothesize that capture bias is more readily externalized and device-normalized (coverage logging, transcription-error rates, held-out sensor comparisons) than self-report bias — while acknowledging that psychometrics has spent decades building corrections for self-report [56,57] and that experience sampling [38] is itself a self-report methodology engineered against retrospective distortion. Comparison arm C tests the hypothesis in both directions, and the self–other asymmetry literature [37] predicts domains where self-report should win.

Digital exhaust is produced *after* the person has already decided what was worth recording; a model that learns only the surface regularities of that residue learns a routine, not a generator. Within passive evidence, digital exhaust (calendars, messages, clicks) is a thin, already-interpreted residue that largely encodes routine, which a strong R2 already captures, whereas higher pre-interpretive tiers can reveal deviations from routine, where skill over R2 is earned; "resolve" and "collapse" are epistemic shorthand for a distribution concentrating, with no physical claim. TargetSpace assumes richer capture is not automatically better: the evidence-tier ablation measures whether passive observation adds lift over digital exhaust and self-report on identical sealed instances (extended basis in Appendices H and G; substrate summary Table 26).

6 The Benchmark Task

TargetSpace is not a model, a dataset, or an architecture. It is a benchmark framework — a task definition, a scoring methodology, standardized baselines, an evaluation grid, and integrity and governance protocols — organized as a shared apparatus with multiple application tracks (Section 8.3). Many datasets may instantiate it and many architectures may compete on it, the sense in which ImageNet is distinct from any network trained on it and SWE-bench [2] from any agent that solves its issues.

The systems under evaluation are **ambient target-state systems**: systems that passively or semi-passively capture longitudinal observations about a person, group, or environment (audio, video, screen traces, location, documents, calendar, interactions, or sensor data) and transform them into a persistent representation capable of retrieving, summarizing, predicting, or steering future-relevant states. "Ambient" does not require continuous recording; repeated passive or semi-passive capture across days to months qualifies. The object of evaluation — and, as Section 13 argues, of governance — is not only the raw signal but the derived memory and model built from it: each layer up from the raw signal (perception, structure, memory, target state) is more compressed, searchable, and predictive than the one below, and TargetSpace evaluates the top of this ladder (Appendix M, Table 27).

6.1 From goals to target states

The object the stack scores is a target system's **target-state transitions**. The intuitive version — predict a system's *goals* — smuggles in conscious intent; we therefore speak of **target states**: configurations a system behaves as if acting to reach and to restore when perturbed, whether or not any mind intends them. Operationally, a target state is a behaviourally resolved configuration the person tends to reach, maintain, resume, or abandon (a commitment, priority, or task regime), scored only through a deterministic resolution rule over a future observable outcome; the mind-agnostic vocabulary lets one apparatus span persons and non-persons without any metaphysical claim [41]. Two clarifications keep the construct honest. First, the scored object is extensional: a scored target state is a pre-registered behavioural event class with a deterministic resolution rule, nothing more; the teleological reading — a configuration the target acts to reach and restore — is an interpretive hypothesis about what generates the regularities, not a property the protocol tests, and perturbation-resumption probes that would test it directly are future work outside the default observational benchmark. Second, the vocabulary has lineage — cybernetic teleology [71], plan structures [72], the intentional stance [73] — and what is different here is only binding it to pre-registered resolution rules, so every use of it is scored. This is also the source of the benchmark's name: at any moment a partially observed target occupies a possibility space — a distribution over the states it might move into next — and an adaptive system does not wander that space at random; it moves toward its target states. TargetSpace scores how well a system predicts that resolution. Transitions matter because the moments when what a system is acting to reach changes are where non-routine behaviour originates and a routine baseline fails, the regime in which target-specific skill is decisive. TargetSpace treats the transition typology (emergence, stabilization, competition, displacement, abandonment, resumption, and constraint- or opportunity-driven change) as a prediction target with deterministic resolution rules (Appendix I), which separates it from recognizing which fixed goal is active (Section J.2). The path a target traces through these transitions is its **observed trajectory**: not a fixed profile of what the target is, but the externally observable sequence of what it is oriented toward, which the benchmark asks a model to anticipate from observable evidence, with no claim about inner experience.

6.2 Definitions and the prediction unit

A **target-system instance** is a specific partially observed adaptive system tracked over time (a person, animal, robot, organization, project, software agent, market). A **latent target state** is a configuration it acts to reach or maintain, and a **transition path** is the sequence of target states it moves through — the object a model is asked to predict. A benchmark instance is the tuple $(\boldsymbol{i}, \boldsymbol{E}_{\leq t}, \boldsymbol{q}, \boldsymbol{A}, \boldsymbol{r})$: instance i; evidence up to time t; a query q about a future state with discrete answer space A; and resolution time $r > t$ with a deterministic **resolution rule**; the system outputs a distribution over A. Queries are organizer-issued, not entrant-chosen, so skill cannot be inflated by predicting only easy instances; and because predictions are nested within instance and serially dependent, inference is performed at the instance level, so the number of independent *instances*, not the raw prediction count, bounds power for population-level claims. Three background conditions are explicit. **A1** states that achievable skill is bounded above [8,9]: it guarantees no positive skill — positive skill over R2 is what the benchmark discovers, never what it assumes — and is one reason no single headline number can represent the construct. **A2** latent target states are evaluated only through observable consequences. **A3** is a measurability precondition rather than an empirical premise: the instance changes, so the benchmark rewards adaptation, and where a target is near-stationary and R2 near-deterministic the benchmark returns an uninformative, not an invalid, result. The measurement is agnostic to which latent structure produces the predictions — the property that makes the grid architecture-neutral (Section 7). Predictions are specified and scored at multiple horizons and abstraction levels — short-horizon concrete next states, medium-horizon routine deviations, and long-horizon abstract transitions — each level scored independently rather than assuming one representation is optimal across horizons (Appendix H, Table 19).

Inclusion rule (hard). *A target state is included in TargetSpace only if it has a pre-registered observable resolution rule; otherwise it is evidence, not a scored state.* Everything else — attention, avoidance, concern, need, stated and inferred goals, and any other latent construct — enters only as **evidence** $E_{\leq t}$ and earns credit only once a pre-registered rule maps a future observable consequence into the answer space. The pipeline is fixed and one-directional — **evidence** → **latent construct** → **prediction question** q → **answer space** A → **deterministic resolution rule** → **scored outcome** — and only the last object is scored (assumption A2). Observable outcomes are not treated as the target-state itself, but as pre-registered consequences through which a latent target-state prediction can be scored. Three corollaries follow. *Attention is evidence*: a continuously observable, hypothesized leading indicator of transitions, never a scored state unless it is itself part of a pre-registered observable outcome (Section 8.2). *Self-report may be evidence but is never the outcome label* for any instance in which it is also a model input, so no system is graded against the very channel it consumed (Section 9.6). *Causal claims* — that some construct *causes* a transition — require intervention or explicit identification and lie outside the default observational benchmark (Section 13). Table 4 illustrates the boundary.

Belief-state framing. One sufficient strategy — not a requirement — is to maintain, from observations $O_{1:t}$, a belief state B_t over latent variables (goals, constraints, salient episodes, recent transitions, and its own uncertainty) and emit the calibrated future distribution it implies. In the POMDP idealization B_t is a sufficient statistic of the history [46]; real targets violate the stationarity that sufficiency requires (A3), so B_t is a useful abstraction rather than a guarantee. TargetSpace does not score B_t: systems with different internal B_t stay comparable because both emit a distribution over the same answer space, and R2 and the permutation gate test whether whatever B_t a system maintains is genuinely about this target.

Several often-conflated notions are kept distinct because the benchmark scores some and not others (definitions in Appendix A): a **stated goal** (what a person declares) and an **inferred goal** (what a model concludes they pursue) are evidence; an **enacted priority** is what allocation of a scarce resource reveals; an **implicit target** is a latent orientation inferred from behaviour, repetition, and avoidance, possibly held without introspective access to it [36]. TargetSpace scores only **predicted future states and their transitions** — a pre-registered discrete state or transition with an externally observable outcome that resolves it (assumption A2) — not stated goals, not attention or affect, and not next-action mimicry, and never a self-report channel the model itself consumed as evidence (Section 9.6): "the person is avoiding this reply" is scored only as, e.g.,

Table 4: Valid versus invalid target states (inclusion rule, Section 6.2). A candidate is a **scored target state** only if a pre-registered rule resolves it against a future observable consequence; otherwise it is **evidence** about a state, or lies outside the default benchmark, and is never scored directly. Each valid candidate is predicted as a probability distribution over its answer space.

Candidate	**Verdict**	**Why**
No substantive reply to message m by r	Valid	discrete answer space; deterministic rule over an observable future action
Commitment c resolves as {complete, defer, cancel, replace} by r	Valid	pre-registered answer space; outcome fixed by calendar/task status
"The person feels anxious about the deadline"	Invalid (evidence)	affect has no observable resolution; scorable only via a mapped outcome, e.g. P(avoidance of obligation by r)
"The person is attending to X now"	Invalid (evidence)	attention is a leading evidence stream, not a scored state, unless part of a pre-registered observable outcome
"The person's true goal is Y"	Invalid (evidence)	inferred goal is a latent construct; scored only through a resolvable future state it predicts
"Attention *causes* the switch to Y"	Invalid (out of scope)	causal claim; requires intervention or identification, outside the default observational benchmark
Self-reported goal G at r, where G is also a model input	Invalid (label)	self-report may be evidence but is never the outcome label when also used as input

P(no substantive response to message m by r).

Protocol elements. Each element of a benchmark instance — target, target state, prediction, observation window, prediction horizon, outcome, ground truth, evaluation metric, baseline, uncertainty reporting, and leakage control — is fixed as a concrete definition with a pre-registered option set before sealing and disclosed with the result, so the protocol is concrete rather than left implicit; the full element-by-element table is Appendix A (Table 15).

6.3 The walk-forward sealed protocol

A retrospective benchmark replays logged histories, so apparent prediction becomes partly retrieval, the failure that motivated contamination-resistant designs [3,4]. TargetSpace privileges a prospective, sealed mode: predictions are timestamped and sealed (SHA-256) before outcomes exist and resolved automatically, under strict walk-forward evaluation (only evidence timestamped $\leq t$, indices rebuilt per slice; random cross-validation prohibited). Sealing binds only when the commitment leaves the submitter's control: the hash must be lodged with an external witness — the organizer's registry or a trusted timestamping service — before resolution, with the clock source declared. A hash generated and held on the submitter's own machine proves nothing; runs without external commitment are labelled self-attested rather than contamination-audited. For sensitive longitudinal first-person data the harness can run where the data lives (federation), exporting only sealed predictions, resolved outcomes, and aggregate reports, so the scoring protocol is separable from the later consent-governed construction of a shared corpus (Section 13). Because target-specific prediction is longitudinal rather than IID the evaluation preserves chronology; and because a prediction is scored against sealed external outcomes rather than internal self-consistency, a system that produces confident rollouts in its own simulator earns no credit until they resolve against what the target actually did

— a guard against a model that is competent only inside its own dream.

6.4 What TargetSpace is not

Not user modeling. A user model predicts a target's *outputs* (clicks, ratings, responses) to tailor a system; a target-state model predicts the *evolving latent state that generates those outputs.* **Not event-outcome forecasting.** Sealed proper-scored prediction of external events is established, including live LLM-forecasting leaderboards over real-world events and regulated prediction markets [13,81,82]; TargetSpace scores transitions in the latent target state of a *tracked instance* with instance-specific controls (R2, permutation) those benchmarks do not use. **Not a chat, retrieval, task-completion, or desktop-automation benchmark.** Those evaluate what a system *does* when asked; TargetSpace evaluates whether a system can infer and predict a person's target states and their transitions from *passive observation up to a sealed time*, before any request is made. **Not a competitor to JEPA** or to latent-prediction research (Section 12): it evaluates such systems rather than replacing them. The shift is clearest by contrast: a transcript benchmark asks *what was said?*; a memory benchmark asks *what happened?*; TargetSpace asks a third question — *what target state is the person in, and what transition is likely next?* The first two score recall of artifacts a person already produced; the third scores prediction of a latent, evolving process.

6.5 Why target-state prediction differs from generation and imitation

Four capabilities are easy to conflate, and the benchmark scores a specific one. **Generation** predicts surface outputs, judged by plausibility; an **imitation or reactive policy** maps observations to actions; **consequence prediction** predicts the future state an action or event produces; **planning** searches over predicted futures to choose actions. These may coexist in one system [24]. Generation is not agency: a system that produces a plausible next action without predicting the state it leads to cannot plan, cannot be held to an outcome, and cannot be distinguished from a confident imitator. TargetSpace scores *consequence prediction* and *target-state prediction*: it does not infer planning ability from action success and requires no human-readable internal simulator — any system that emits a calibrated distribution over externally resolvable future states may participate, whether a reactive policy, a latent world model, an LLM, a symbolic planner, or a hybrid. It also does not ask a system to predict every ripple in the river — every pixel or token, much of which is irrelevant or unpredictable — but whether it identifies the *target-relevant* transition, the same intuition that motivates predicting in a learned representation rather than reconstructing raw observations (Section J.1), lifted to the measurement layer. An optional action-conditioned mode makes consequence prediction explicit while keeping passive prediction, observational counterfactual prediction, actual intervention, and causal-effect estimation distinct: a prediction scored on observational data is not a causal estimate, since causal identification requires intervention or explicit structural assumptions (Section 13).

7 The TargetSpace Validity Stack

We present the controls not as a loose conjunction of desirable properties but as a stack: each layer removes a way a prediction can look like understanding without being target-specific. The core of the stack — layers 1–7, the *target-specificity stack* — is mandatory; extended layers (Section 7.1) and deployment levels (Section 7.2) define how much of the full apparatus a given evaluation carries. TargetSpace adopts most of these tools from prior work (Section 12); its contribution is the conjunction, assembling them around a single question, whether a prediction is genuinely about the target system, with the own-routine baseline and the permutation gate as the two layers that make the stack test target-specificity rather than generic predictive quality.

Table 5: The target-specificity stack. Layers 1-3, 5, and 7 are established tools we adopt and cite; layers 4 and 6 — the own-routine baseline and the permutation specificity gate — are the anchors that make the stack a test of target-specificity rather than of generic predictive quality. Most components are adopted from prior work; the novelty is their assembly around one measurable question, not isolated invention of any part.

Layer	What it removes / certifies	Role in TargetSpace
1. Prospective sealing	predicts sealed and timestamped before outcomes exist — removes hindsight rationalization and contamination	adopted tool [13]
2. Proper scoring	log / Brier scoring of the whole distribution — rewards calibrated probabilities, not cherry-picked accuracy	adopted tool [10–12]
3. R1 population-prior baseline	skill must exceed population base rates — filters out generic base-rate prediction	adopted tool [48]
4. R2 own-routine baseline	skill must exceed the target's *own* strong routine — filters out routine mimicry	**TargetSpace anchor**
5. Calibration gate	calibration intercept/slope and reliability (ECE a coarse diagnostic at small n) — filters out overconfident lucky prediction	adopted tool [6]
6. Permutation specificity gate	skill must collapse when predictions are matched to the wrong target — tests dependence on the correct instance pairing	**TargetSpace anchor**
7. Evidence ablation	skill as a function of evidence tier — measures which streams add target-specific information	adopted tool [39,40]

The stack composes into a single decision rule: a prediction earns target-specific credit only if it beats R1, beats an admitted R2 (one that itself beats R1), passes calibration, and fails under permutation. Removing any one layer reopens a way to score well without target-specific skill, which is why we present them together rather than as separable contributions. We are equally explicit about what a passed stack certifies: calibrated, person-specific predictive skill beyond R2's declared feature set. It does not, by itself, certify dynamic latent-state modeling: a stable idiosyncratic rule outside R2's features — this person cancels whatever collides with a standing obligation — beats R2 and collapses under permutation while modeling no dynamics at all. Any claim about transition modeling therefore rests on the routine/transition stratification (Section 9.6), where skill concentrated in transition cases is the evidence that separates dynamics from a static per-person profile. Beating R2 is the line between modeling the target's routine and modeling the target; the strata say which kind of modeling it was.

A hierarchy of claims. What a result licenses is graded, and the apparatus certifies each level separately rather than letting one leak into the next. **Level 1 — task skill**: the system predicts one behavioural outcome above R1. **Level 2 — target-specific task skill**: it beats an admitted R2, stays calibrated, and its skill collapses under wrong-target permutation. **Level 3 — dynamic target modeling**: its advantage depends on correct temporal order (the shuffled-history control) and concentrates in transition cases. **Level 4 — reusable personal representation**: a frozen representation transfers prospectively to held-out task families without labeled adaptation (Section 8.4, hypothesis H6). **Level 5 — downstream utility**: the validated representation improves assistance under a separate intervention protocol (hypothesis H7). The core benchmark certifies

Levels 1–3. Levels 1–2 alone are *evidence consistent with a target-specific personal representation*, never a demonstrated personal world model; describing a particular system as maintaining a personal world model is language earned only at Level 4, across multiple, diverse held-out families. This grading is the paper's answer to a fair objection — that a battery of task predictions might measure a collection of narrow behavioural predictors rather than a model of the person: the objection is correct about Levels 1–2, addressed by temporal-order and transition evidence at Level 3, and settled only by transfer at Level 4.

Holding the prediction unit and scoring fixed, the stack is applied across an evaluation grid that varies *what a system may observe* (the evidence tier, from low-content behavioural metadata, L0, to physiological sensors, L6) and *what kind of system it is* (the architecture class: LLM, VLM, multimodal agent, latent-predictive, symbolic/probabilistic, reactive policy, hybrid, plus human self-report and oracle anchors). Every cell is comparable because every system emits a distribution over the same answer space on the same sealed instance, through a disclosed output adapter reported as part of the system under test. The full architecture-class and evidence-ladder specifications, including the per-tier privacy-risk taxonomy, are in Appendix L; whether richer evidence helps is measured by the tier ablation (skill over R2 per tier), never assumed.

7.1 Extended validity layers

Beyond the core, six layers extend what a passed evaluation can license. **(8) Retrieval-only comparison.** The same evidence is run through a retrieval-and-summarization pipeline with no predictive model; lift the system cannot show over this arm is lookup, not modeling. **(9) Regime-shift and novelty tests.** The routine/transition stratification (Section 9.6) is the novelty test: instances where the target departs from ordinary routine are scored separately, because shallow repetition suffices everywhere else. **(10) Out-of-distribution degradation curves.** Instances are ordered by a pre-registered distance-from-routine measure (for example, R2's own surprisal on the realized outcome), and skill is reported as a curve over that distance rather than a single average: structural target modeling should degrade gracefully as situations become less routine, where shortcut prediction collapses. **(11) Counterfactual probes (advanced).** The system estimates how target states would change under altered conditions or constraints; scored only where natural experiments, pre-registered condition changes, or domain structure provide verifiable ground truth, and treated as partly domain-specific [85]. **(12) Explanation faithfulness.** Where a system offers explanations for its predictions, the cited evidence must match the evidence that actually drives them: masking the evidence an explanation names should degrade the prediction accordingly, so explanations cannot merely rationalize outputs after the fact. **(13) Internal-representation probes (optional).** Where model internals are available, probes test whether hidden states encode stable target variables, temporal structure, uncertainty, and evidence dependencies; valuable, but never required, because many systems under evaluation are black-box. TargetSpace does not ask only: did the model predict the target's future? It asks: did it predict for the right reasons, under uncertainty, under perturbation, and beyond routine?

7.2 Deployment levels

The full apparatus defines the research direction; early implementations can start smaller without forfeiting validity. Three levels fix what a reported evaluation must include.

Table 6: Deployment levels. Each level is a strict superset of the one before it; a report states its level, and claims are licensed accordingly.

Level	**Required components**
Minimal	prospective state-transition prediction; temporal sealing and leakage control; calibration; generic (R1) and own-routine (R2) baselines
Standard	Minimal, plus wrong-target controls; retrieval-only comparison; evidence ablation; regime-shift / novelty stratification; OOD degradation curves; evidence-cost reporting
Gold-Standard	Standard, plus counterfactual probes; explanation faithfulness; full privacy-efficiency reporting; optional intervention modules where ethical and practical; optional internal-representation probes where internals are available

8 Personal-Track Instantiation

The personal track is the flagship instantiation and the only one this paper carries beyond formulation, because understanding individual people is valuable and the observation bottleneck bites hardest there. Here the target is most vividly an **observed trajectory** (Section 6.1), and the evidence tiers are its observable traces. The person-domain **evidence ladder** (Appendix L, Table 25) runs from already-externalized *digital exhaust* — calendar and communications metadata (**L0**) and user-authored text and transcripts (**L1**) — through *passive audio* (**L2**), *screen and egocentric audiovisual capture* (**L3–L4**), *location and mobility* (**L5**), to *physiology and, prospectively, affective or internal-state proxies* (**L6**). Higher tiers may offer more independent access to internal state, but the benchmark still scores only *observable future outcomes*; it never accesses the person's experience itself. The target-system instance is a consenting individual; the scored target states are categories such as commitment active versus abandoned, priority maintained versus displaced, and task continuation versus switch (Appendix I), while attention, concerns, and avoidance are *evidence* about those states rather than states themselves. The personal-AI question is sharpened by target-specificity: not merely 'does passive capture help?' but 'does passive first-person evidence help a model predict *this person's* future target-state transitions beyond population priors, beyond this person's own routine, and beyond their own self-report?'

Concretely, the **prediction targets** are observable future states and transitions — commitment follow-through, meeting and event realization, response behaviour, task continuation versus switch, priority maintenance versus displacement, and engagement versus avoidance of a defined obligation — each with a deterministic resolution rule (Table 20); attention allocation, social context, and emotional cues are **evidence**, and deadlines, dependencies, and competing obligations are **constraints**. The distinction is concrete at the level of individual traces: a location ping is not intent, a task checkbox is not a priority, and a transcript is not a meeting's latent importance. Each is an observation that becomes scorable only once a pre-registered rule maps it to a future observable resolution — whether the obligation is engaged or avoided, the reply eventually sent, the commitment completed, deferred, cancelled, or replaced. The evidence hypothesis is the one stated in Section 5: exhaust encodes routine, and higher tiers may reveal the deviations — the shifted attention, the commitment about to slip — where skill over R2 is earned. The hard case is accumulation: individually banal observations become predictive when combined across time, which is why the benchmark evaluates inference from longitudinal traces rather than question answering over transcripts. Audio-only traces are the early form of this evidence [49–52]; audiovisual first-person traces are a candidate stronger future form because they add embodied context — what the person sees, where they are, which objects, documents, and screens are present, who is nearby, and what actions are physically taken [34,35,53–55]. The move from audio to audiovisual first-person capture expands the TargetSpace problem from conversational memory to embodied experience modeling; in the evidence ladder both are simply tiers (L2 versus L3–L4), and whether richer capture pays is measured by the ablation, never assumed.

8.1 A minimal runnable task

To fix ideas, one pilot-ready task — the *Recurring-Commitment Completion Prediction* — is runnable on current language models, agents, and retrieval or memory systems. Given timestamped evidence up to a sealed time T (calendar and task history, communication metadata, prior completion/defer/cancel history; optionally text traces and passive-observation summaries), the system predicts how a scheduled recurring commitment resolves by resolution time r, over the answer space $A = \{\text{complete, defer, cancel, replace}\}$. R1 and R2 are fit walk-forward on evidence before T; a deterministic rule over later observable evidence (calendar status, attendance, task completion, or a pre-registered behavioural marker) fixes the outcome; scoring is log score (bits) and Brier against the sealed outcome, reported as skill over R1 and over R2; and a permutation test across matched targets must reduce or collapse the skill. The full specification is Appendix H. This is the shape of every personal-track task: the transition types it draws from are catalogued in Appendix I. For a builder, it makes a memory or context feature auditable — run the same sealed task with and without the feature and read its prospective lift over R1 and R2, across evidence tiers and architectures (Table 2); a feature that fails the comparison may still be valuable retrieval or personalization, but it has not shown that it models the user's evolving target state.

8.2 Attention-conditioned predictive lift

In the personal track, attention allocation is the leading evidence stream: it reveals what competes for a person's limited resources [42,43] and is a continuously observable, hypothesized leading indicator of transitions, evidence of allocation rather than a read-out of goal identity (in non-person domains the analogue is whatever scarce resource the instance allocates). **Attention-conditioned predictive lift** is the difference in target-specific Skill between a model conditioned on the attention trajectory and a matched model that omits it, measured beyond R1, R2, stated goals, and contemporaneous behaviour, with the attention proxy, granularity, missingness, and leakage controls pre-registered and lift scored only on sealed outcomes. A positive lift demonstrates incremental predictive value, not causation: attention may proxy an unobserved cause, and causal identification would require intervention (Section 13). Whether attention also helps *form* the next target state is a falsifiable ablation hypothesis, not the core contribution; active inference is one compatible framing [16], and the extended treatment is in Appendix G (Section 5).

8.3 The multi-track apparatus

TargetSpace is a shared apparatus, not a single dataset: **personal intelligence** (TS-Personal) is the flagship track, and additional tracks — health, energy, robotics, enterprise — are specified in Appendix L only to show the framework is not merely a personalized-assistant benchmark. Each track preserves the common scoring spine of Section 7 [6] and is admitted only when it requires a *distinct* validator, evidence band, and horizon profile **and** admits a strong R2; otherwise it is a regime within an existing track, or not yet a track (the admission rule, and why one-shot cross-sectional problems such as cell-fate are excluded, are in Appendix L). We report no results for any track other than the synthetic personal instantiation, and make no cross-domain empirical claim.

8.4 Cross-task personal transfer (advanced track)

The target-specificity stack certifies skill on the task it scores, but a single task cannot distinguish a reusable model of the person from a narrow rule tuned to one outcome. The stronger claim — that a system has learned a *representation of the person* rather than a per-task heuristic — calls for a transfer test. We define one, as an advanced track and explicitly not a requirement of the feasibility pilot.

Definition. A personal representation demonstrates **cross-task transfer** when it improves sealed Skill over R2 on a *held-out target family* without being updated using labeled outcomes from that family. The representation is built from a disclosed set of source task families and frozen at a declared time; the held-out family is scored under the same sealing, resolution, and gating rules, with no labeled adaptation on it.

Candidate held-out families, any of which can be withheld while the others build the representation, include response timing, commitment resolution, event realization, task continuation versus switching, priority maintenance versus displacement, and engagement versus avoidance of a defined obligation. The specification fixes: the source (representation-building) evidence; the held-out task family; the constraint that no labeled outcomes from that family were used for adaptation; identical sealing and proper scoring; the R1, R2, calibration, and permutation controls applied to the held-out family; a comparison against a task-specific baseline trained directly on that family (transfer is credited only when the frozen shared representation approaches or exceeds it); and target-clustered uncertainty, since the inferential unit remains the person. Transfer that survives these controls is evidence that the model carries person-level structure reusable across the ways that person's target states resolve — the property a per-task rule lacks. Three further rules keep the claim honest. Permitted adaptation is declared: unsupervised updates from the held-out period's *evidence* stream may be allowed under the declared adaptation policy, but any use of the held-out family's outcome labels voids the transfer claim. Failure is graded: held-out skill that vanishes indicates a per-task rule; skill that survives but falls far below the task-specific baseline indicates partial, weakly reusable structure — both are reportable results. And the language is earned incrementally: describing a system as maintaining a *personal world model* requires transfer across at least two held-out families from distinct behavioural domains (e.g., response timing and commitment resolution), each passing the full gate set; one family is a data point, not a model. This is hypothesis H6 (Section L.2); it is future work, and no transfer result is claimed here.

9 Evaluation Protocol and Pilot Plan

The scoring foundation is established practice we **adopt and cite**; we then add metrics for possibility-space resolution. 'Collapse' is the flagged epistemic metaphor of Section 5 — a distribution concentrating — with no physical meaning.

9.1 Preserved foundation (adopted)

Predictions are scored with strictly proper rules: the **log score** (primary; the log-likelihood $\ln p$ of the sealed outcome, so higher is better) and **Brier score** (secondary) [10–12]. The headline quantity is **Skill**, in bits:

$$Skill = \frac{\text{mean log-score}_{\text{system}} - \text{mean log-score}_{\text{reference}}}{\ln 2},$$

a paired comparison on identical sealed instances. In expectation Skill equals a difference of divergences from the truth, $D(P\|\text{ref}) - D(P\|\text{sys})$, and reduces to a mutual-information gain only in the special case where the reference matches the true marginal and the system the true conditional [11]; since R1 and R2 are estimated, we read Skill as relative log-score improvement in bits — a plug-in estimate whose reference error propagates into the bit count — not as a measured information quantity. Skill is reported against **R1** (population prior, estimated leave-one-out without the scored target's history, matched on question type, answer space, and horizon; the entry condition) and **R2** (the target's recency-weighted, walk-forward routine; the target-specific condition). R2 is pre-registered, organizer-fit, and frozen (Appendix D); it is admitted only if it beats R1 on historical walk-forward slices for the same question family, answer space, and horizon, and otherwise that task is not counted as a target-specificity test — so 'skill over R2' cannot be manufactured against a weak baseline. **Calibration** gates the result [6] primarily through the calibration intercept/slope and a reliability decomposition, with top-label ECE bands ($\leq$0.10 pass / $\leq$0.20 warn / $>$0.20 fail) as a coarse secondary diagnostic; for multiclass answer spaces the intercept/slope diagnostics apply one-versus-rest per class (classwise calibration [75]), classes below a pre-registered minimum event count are reported as non-assessable rather than passed, and a top-label diagnostic never stands in for full-distribution calibration; ECE's small-sample bias and the pilot-scale treatment of the bands as diagnostic bounds are specified in Appendix K. The **permutation specificity gate** scores a system's model for instance i against another instance's outcomes, matched within compatible question types, horizons, and answer spaces so target identity is not confounded with base rates; skill that does not collapse was not target-specific. The default report compares true-instance Skill over R2 with matched-permutation Skill over R2: a pass requires the true-instance estimate to exceed the matched-permutation estimate by a pre-registered margin, with uncertainty reported. The pass criterion is the margin exceedance, never the mere absence of permuted skill — low power makes apparent collapse easier to observe, not harder, so an underpowered run cannot back into a specificity certificate — and at pilot scale the result is diagnostic rather than confirmatory (Appendix K). What a passed gate licenses is instance-specificity: dependence on the correct prediction-to-outcome pairing, no more. Of this stack, R1, proper scoring, calibration, and sealing are adopted directly from forecasting practice [6,10–13]; the R2 baseline and the permutation gate are standard tools placed in a role that is not standard — target-specificity control — so the contribution is their conjunction, not any element in isolation. Full reporting details — probability floors, stratified aggregation, recalibration policy, permutation-effect reporting and power, and day-blocked / person-clustered uncertainty — are in Appendix K.

9.2 A worked instance (synthetic, illustrative)

To make the apparatus concrete, we work through one synthetic instance (invented numbers, no participant, no result claimed). A single tracked person has a recurring Wednesday review on the calendar, but this week passive signals show attention redirected to an urgent dependency. The query is whether the review is completed by Wednesday 17:00 (answer space {completes, defers}). R1 (population prior), R2 (this person's own routine), and the evaluated model assign $P(\text{defers})$ of 0.15, 0.10, and 0.70; the review is deferred, so the model gains about +2.2 bits over R1 and +2.8 over R2 — skill the routine did not contain. Scored against a *different* person who kept their review, the same prediction earns about −1.6 bits over that person's R2, so its skill collapses and is confirmed specific to this target. A model that had learned only base rates or this person's habit would match R1 or R2 and show no lift. (All numbers are illustrative.)

9.3 The control battery

TargetSpace is not only a dataset; it is an **evaluation protocol** for whether longitudinal target information improves prediction in a measurable way. A meaningful evaluation runs a model across a battery of conditions, each instantiating machinery defined earlier: a *zero-history* arm (operationally R1), a *short-history* arm (locating where added history first pays), and a *longitudinal-history* arm (the condition the benchmark rewards); a *shuffled-history* control (the correct target's history in scrambled temporal order — a re-scored diagnostic that never alters a sealed prediction and isolates *order*) and a *wrong-target* control (the permutation gate, isolating *identity*); *ablated-modality* controls (the evidence-tier ablation); *oracle* and *human* anchors bounding the achievable range; *calibration*; and *evidence attribution* (the model identifies the observations supporting a prediction, enabling the explanatory audit a passed gate licenses). The three history arms plus the wrong-target gate suffice to substantiate the core claim; the remainder are recommended diagnostics. The shuffled-history and wrong-target controls separate the two ways a prediction can be target-specific in name only: indifference to *when* the evidence arrived, and indifference to *whose* evidence it was.

Ablation logic. The benchmark's core claim is validated through these ablations, not asserted. *The benchmark's central signal is not merely whether a model predicts the future, but whether prediction improves when the model is given the correct target's history in the correct temporal order.* If target history, temporal order, target identity, or multimodal evidence do not improve performance, the benchmark is not measuring its intended capability, and the result should be read as null; improvement under the true conditions that degrades under the shuffled-history and wrong-target controls is the inference the stack licenses (Section 9).

The battery pairs with a fixed set of baselines (trivial-prior R1, recency, single-modality L0/L1/L2, multimodal, long-context, and human where feasible; a **retrieval-only arm** — the same evidence through a retrieval-and-summarization pipeline with no predictive model — is required at Standard level and above, since lift a system cannot show over it is lookup, not modeling) and ablation dimensions (observation-window length, modality removal, target permutation, memory/retrieval access, prediction horizon, and temporal-leakage controls), each reported per evidence tier and per horizon and never pooled into a single headline number; the full specification is in Appendix K. Table 7 consolidates how each characteristic partial-pass pattern across these baselines and gates is read: a result counts as genuine target-specific signal only when it clears every row at once.

9.4 A research-hypothesis ladder

The controls above are not a checklist of equals; they form a ladder of progressively stronger, separately falsifiable claims (Table 8). Each rung presupposes the ones below it, and a program can stall or fail at any rung — a failure that is itself an informative result (Section 14). The ladder also disciplines what this paper currently supports: the synthetic harness exercises the *mechanics* of H0–H4; the feasibility pilot *examines* them without the power to confirm them; H5–H7 are later or future work, claimed here only as method. Each rung is assigned to a phase of the staged roadmap (Section 9.7).

9.5 Reporting contract and protocol card

Table 9 states the whole protocol as a card a team can implement now.

TargetSpace is meant to be run, not only read. A researcher choosing to evaluate a system specifies, in order: **(1)** the target-system instances; **(2)** the prediction questions and their deterministic resolution rules; **(3)** the evidence tiers the system may access; **(4)** the architecture class(es) under test; **(5)** the R1 population-prior and R2 own-routine baselines; **(6)** the sealed

Table 7: Failure-mode read-out: each observed pattern maps to the diagnosis it licenses. A result is genuine target-specific signal only when it clears every row at once; any single pattern below downgrades the claim.

Observed pattern	**Diagnosis (claim downgraded to)**
Beats R1 but not R2	*Routine replay*: recovers population base rates and the target's own habit, with no skill beyond memory.
Beats R2 but fails calibration	*Overconfident or lucky*: the bit gain is not trustworthy; recalibrate or abstain before any claim is made.
Skill survives the permutation gate	*Not target-specific*: skill rides base rates or structure shared across targets, not this target's dynamics.
Lift appears only under richer capture with poor coverage	*Capture artifact*: the gain tracks instrumentation and coverage, not target state; check coverage logs and held-out sensor comparisons.
Improves only on routine-continuation cases, not transitions	*Not transition modeling*: predicts habit persistence, not target-state change — the capability the benchmark most rewards.

prediction times and horizons; **(7)** the scoring metrics; and **(8)** the permutation test. The federated harness then runs where the data lives and reports a standardized row: **Skill vs R1**, **Skill vs R2**, calibration (pass/warn/fail), **permutation result** (collapses / survives), **predictive lift by evidence tier**, the **number of resolved predictions** *and* the **number of independent targets** (the inferential unit for between-person claims — a large prediction count from few targets is not a large sample), the **horizon**, the **target domain**, and the **minimal sufficient tier** — the lowest evidence tier whose gated skill falls within a pre-registered equivalence margin of the best arm, the field that makes evidence minimization a ranking rule rather than a preference. The reporting contract is the point: every row discloses, simultaneously, whether the system beat the average, beat the routine, stayed calibrated, and depended on the correct target — so a reader can tell target-specific predictive skill from generic prediction at a glance. Comparability has a trust model: because Skill over R2 is cohort- and baseline-relative, and because a self-run federated evaluation makes the entrant simultaneously cohort-selector, R2-fitter, and outcome-resolver, self-run outputs are self-attested within-study diagnostics, not leaderboard entries. Cross-system ranking requires organizer-issued queries, an organizer-fit frozen R2, third-party or rule-deterministic outcome resolution, and externally committed seals — the distinction the verification-level field of every row records (Appendix B).

9.6 Task tracks, pilot plan, and status

Version 1 uses auto-scorable, high-frequency targets organized into four domain-neutral task families instantiated for persons: **TS-R** short-horizon realization (next contact; event realization; response behaviour); **TS-A** attention/allocation; **TS-D** decision/commitment (commitment follow-through); **TS-X** target-state transition/drift.

The pilot's experimental question, stated sharply, is whether passive longitudinal observation improves calibrated, target-specific prediction over each of five references: (A) the population prior R1; (B) digital exhaust; (C) self-report and stated goals, where available; (D) the person's own routine R2; and (E) the wrong-target permutation; Table 10 states what beating each reference demonstrates. This is a *feasibility* study only; the confirmatory test of these comparisons is a later, powered study (Section 9.7).

Table 8: The research-hypothesis ladder. Each rung is a separately falsifiable claim presupposing those below it; the right column keys each to a phase of the staged roadmap (Section 9.7). Nothing above mechanics is claimed as an empirical result in this pre-pilot paper.

Rung	**Claim**	**Where addressed**
H0	Generic predictability: outcomes are predictable above population priors (R1).	Phase 0 mechanics; Phase 1 examines
H1	Routine predictability: the target's own history improves prediction over R1 (an admitted R2).	Phase 0 mechanics; Phase 1 examines
H2	Target-specific information: the system beats R2 and loses skill under wrong-target permutation.	Phase 1 examines; Phase 2 confirms
H3	Longitudinal dynamics: correct temporal order beats shuffled history, especially on transition cases.	Phase 1 examines; Phase 2 confirms
H4	Observation value: added passive or multimodal evidence beats digital exhaust and self-report.	Phase 1 examines; Phase 3 confirms
H5	Evidence efficiency: comparable gated skill at less energy, data, burden, capture, or privacy exposure (minimum sufficient observation).	Phase 3 (Section 10.4)
H6	Cross-task personal transfer: a shared personal representation improves sealed predictions on held-out target families without labeled outcomes from them.	Phase 4 (Section 8.4)
H7	Downstream utility: a validated personal model improves assistance under a separate intervention, safety, and deployment protocol.	Phase 5 (Section 13)

The first pilot is harness validation: a feasibility study of whether the protocol can run (prediction generation, sealing, deterministic resolution, variance and dependence estimation, and preliminary signal over R1/R2) — Phase 1 of the roadmap, with Phase 1's license only. Concretely, it will be an audio-first exploratory study: five consenting participants, thirty days, sealed daily calibrated predictions, systems differing only in evidence (A population prior; B digital exhaust; C +text/transcript and available self-report; D +continuous first-person stream; model and prompts fixed across B–D), with a human self-report baseline where feasible. The primary endpoint is the paired prospective log-score improvement over R2, $\Delta L(M, \mathrm{R2}) = L(M) - L(\mathrm{R2})$ in bits per

Table 9: The TargetSpace protocol card. Every element is pre-registered before sealing and disclosed with the result.

Element	Specification
Target definition	a person, team, organization, patient, project, device, or environment, tracked as a persistent instance with consistent identity
Observation window	the period during which evidence is collected, with disclosed coverage and missingness
Prediction window	the future period whose state transitions are predicted; horizons pre-registered
Target state variables	pre-registered discrete state or transition classes with deterministic resolution rules
Evidence streams	the information the system may access, by tier, under strict walk-forward access
Baselines	generic prior (R1); routine/autocorrelation (R2, admitted only if it beats R1); retrieval-only over the same evidence; wrong-target substitution
Sealing	predictions timestamped and hashed before outcomes exist; external witness for contamination-audited runs
Scoring	log score (bits, primary) and Brier (secondary); calibration intercept/slope and reliability; lift over each baseline
Reporting	evidence used and its cost vector (volume, sensitivity, collection burden, energy, latency, consent burden, privacy risk); ablation results; independent-target count; deployment level

Table 10: Pilot comparisons and what beating each one demonstrates. A feasibility check of these comparisons, not a powered test; self-report is a comparison and evidence channel, never the outcome label.

Reference	What it supplies / holds fixed	What success demonstrates
A. Population prior (R1)	population base rates; no target-specific information	skill exceeds base rates (entry condition)
B. Digital exhaust (L0–L1)	already-externalized calendar, metadata, and text	passive capture adds beyond a routine-heavy residue
C. Self-report / stated goals	the person's own declarations, where available	observation adds beyond what the person says
D. Own-routine (R2)	the target's recency-weighted, walk-forward routine	skill exceeds habit and memory replay (headline)
E. Wrong-target permutation	predictions scored against another person's outcomes	skill is specific to this person (must collapse)

prediction (positive when the system beats R2), with within-person day-blocked and between-person person-clustered uncertainty; performance over R1, calibration, and skill loss under permutation are validity gates. With five participants the study is not powered to confirm population-level effects: a near-deterministic R2 leaves little headroom and the permutation null is coarse, so it supplies inputs for simulation-based power planning of a later confirmatory study rather than a confirmatory result. Abandonment criteria: no system beats R1 (noise); none beats R2 (target-specific prediction not demonstrated); skill survives permutation (not target-specific, read as a directional flag at this cohort size).

Pre-registration and controls. Prediction instances are organizer-issued under a pre-registered eligibility and sampling frame — not entrant- or retrospectively selected — stratified across routine-continuation and transition cases, with skill reported separately on each, since skill concentrated in transitions is stronger evidence of structure beyond routine. Outcomes follow pre-registered rules (deterministic behavioural outcomes preferred; ambiguous classes use prediction-blind adjudication with reported inter-rater reliability), and an outcome label is never the same self-report channel a model uses as evidence. Negative controls comprise target permutation, a leakage/known-null canary, and a temporal-specificity check. Before any data collection the protocol pre-registers R1/R2 fitting, sampling frame, answer spaces, resolution rules, calibration split, permutation matching, missingness and abstention handling, primary endpoint, multiplicity rule, and analysis; the full specification is Appendix K.

9.7 The staged roadmap

The program is staged so that each phase asks one question, produces one artifact, and licenses one claim — and no phase borrows the license of a later one. **Phase 0 — synthetic harness (complete).** Question: is the pipeline correct? Artifact: the reference harness and schemas. Success licenses mechanics only; failure is a bug, not a finding; it establishes nothing about people. **Phase 1 — feasibility pilot (next).** Question: does the protocol run end-to-end on real, consented lives — sealing, resolution, missingness, variance and dependence estimates, R1/R2 headroom? Artifact: a completed five-person, thirty-day run with abandonment criteria evaluated. Success licenses feasibility statements and power inputs, never effect claims; failure redirects task design before any scale-up. **Phase 2 — powered target-specificity.** Question: do H2–H3 hold at a sample size set by simulation from Phase-1 estimates? Success licenses the core target-specific claim; a null bounds what current systems extract from the tested evidence — a publishable finding, not a dead end. **Phase 3 — evidence-tier and frontier study.** Question: which observations pay, at what cost (H4–H5)? Artifact: measured observation-value curves, the model–evidence frontier, and minimum sufficient tiers. **Phase 4 — cross-task transfer.** Question: is the representation reusable (H6, Section 8.4)? **Phase 5 — downstream utility.** Question: does validated prediction improve assistance (H7), under a separate intervention, safety, and deployment protocol? Sample sizes for Phases 2–5 are deliberately not stated: they are outputs of Phase 1's variance estimates, not promises made before them.

9.8 Release roadmap

The benchmark itself is versioned separately from the research phases, and this paper is explicit about which release it is. **Version 0 (this preprint).** Defines the benchmark, terminology, protocol card, validity stack, deployment levels, and reporting standard; includes example task schemas and the pilot design; ships a synthetic reference harness. It claims no large-scale validation, no dataset, and no empirical leaderboard — its contribution is the protocol, the benchmark specification, and the evaluation standard. **Version 1.** Publishes the first pilot-derived benchmark tasks and baseline results, with a DOI-linked benchmark release (corresponding to Phases 1–2). **Version 2.** Expands targets, modalities, and task families; adds public leaderboards and certified evaluations (Phases 3–5). Nothing in this paper should be read as claiming that a dataset or empirical leaderboard already exists.

10 Hardware and Observation Architecture for Personal AI

The benchmark turns observation hardware into a measurable variable. Every device that captures longitudinal evidence occupies a point on two axes: what it observes (the evidence tiers of Section 8) and what that observation costs in energy, burden, and risk. This section states the research program that follows. It is formulation, not product guidance; it makes no hardware claims and reports no measurements.

We call the program *observation science*: the disciplined study of which evidence streams are needed to model a target over time, how costly they are to collect — financially, computationally, socially, ethically, and energetically — and how much predictive lift they provide. TargetSpace is not only a benchmark for models. It is also a benchmark for observation: which evidence is worth collecting, at what cost, and for how much target-specific lift? Text logs, calendars, messages, location traces, wearable streams, app events, team communication, and environmental sensors all have different signal-to-cost profiles, and the benchmark deliberately does not reward the systems that simply collect the most data. The name proposes a research program, not an established field. It draws on established fields — personal sensing, lifelogging, mobile and ubiquitous sensing, egocentric vision, activity recognition, wearable computing, privacy-preserving learning — and invents neither passive sensing nor personal modeling: person-dependent evaluation, personalized-versus-population comparisons, and sensor ablations are established sensing practice — from continuous multi-month smartphone and wearable sensing of individuals [65,66,80] — and the contribution here is to standardize and harden them — sealing, proper scoring, specificity gates, disclosed cost — into one comparable measurement. The contribution is the measurement link: TargetSpace connects an observation choice to the gated, target-specific predictive skill it produces, so that questions about sensing can be answered by experiment rather than assumption. The central question decomposes into the themes of Table 11: which modalities matter, how much coverage is required, when additional sensing stops helping, which evidence improves skill beyond the target's own routine, how missingness should be represented, and how acquisition should be optimized under energy, burden, and privacy constraints — and, on the model side, which architectures best convert the same evidence into skill, a question the architecture-neutral grid (Section 7) already carries.

10.1 Three device classes

Scientific instruments are existing devices used for Layer 1 validation: phones, audio recorders, wearables, and glasses, selected for evidence coverage, exportability, and privacy architecture rather than product features. The pilot of Section 9.6 requires nothing more than instruments that already ship. Observation devices are future hardware optimized for a single objective: maximize calibrated, target-specific evidence per unit cost, under consent. No shipping product is optimized for this today; current wearables appear to allocate most of their budgets to interaction. Interaction devices are the consumer products of Layer 3 — assistants, displays, glasses with output — which consume a validated model rather than produce one. These classes are roles, not product categories. One physical device may play more than one role, but the objectives conflict: an interaction surface invites exactly the reactivity an observation instrument must minimize, so the roles should be designed, and evaluated, separately.

10.2 Evidence efficiency

The evidence-tier ablation (Section 7) measures what each observation stream adds. Hardware design needs the same quantity normalized by cost. **Evidence efficiency** is a family of experimental ratios, not a single score: gated, target-specific predictive lift per disclosed unit of observation cost. The general form is conditional and incremental,

$$\mathrm{EE}_c(E_k \mid E_{\mathrm{base}}) \;=\; \frac{\mathit{Skill}_{\mathrm{R2}}(M;\, E_{\mathrm{base}} \cup E_k) \;-\; \mathit{Skill}_{\mathrm{R2}}(M;\, E_{\mathrm{base}})}{\mathrm{Cost}_c(E_k)},$$

where E_{base} is a pre-registered baseline evidence configuration, E_k an added modality, sensor, or sampling policy, $\mathit{Skill}_{\mathrm{R2}}$ prospective skill over the own-routine baseline in bits, and Cost_c a disclosed observation cost with a pre-registered boundary. The numerator is admissible only when the run passes every gate: prospective sealing, an admitted R2, acceptable calibration, collapse under wrong-target permutation, disclosed coverage, matched model architecture across arms, pre-registered evidence boundaries, and deterministic outcome resolution. Efficiency computed from ungated skill is descriptive only and is never a target-specific claim. EE is also conditional by construction: it measures the marginal value of E_k to a disclosed architecture at a disclosed baseline configuration, not an architecture-free property of the evidence — a null increment may indict the model's ability to use a stream rather than the stream itself; increments are non-additive and ordering-dependent; negative increments are admissible results. Cross-study comparison therefore fixes the reference E_{base}, the arm ordering, and the denominator unit as part of pre-registration.

Five cost instantiations are defined. *EE-energy*: incremental skill per joule, with the default boundary set at capture energy plus required edge preprocessing; transfer, storage, training, and inference energy are reported separately, and an end-to-end figure may be reported in addition, never instead. *EE-wear*: incremental skill per worn hour — whether a device extracts useful evidence from the time a person actually tolerates wearing it. *EE-capture*: incremental skill per valid capture hour, separating wearability failure from sensor productivity. *EE-data*: incremental skill per captured bit, an information-exposure and storage-efficiency measure. *EE-interaction*: incremental skill per required manual activation, a burden measure. Privacy is deliberately not collapsed into a scalar: it is reported as a vector — raw audio hours, image count, video duration, bystander exposure, retention period, identifiability, processing locality, derived-inference retention — and treated as a Pareto dimension (Section 10.5). The full evidence-cost family reported alongside any lift claim comprises predictive lift, evidence volume, evidence sensitivity, collection burden, compute and energy cost, latency, user consent burden, privacy risk, and the marginal value of each evidence stream. We deliberately do not name any of these "understanding per joule": understanding is not the measured object; gated predictive skill is. Evidence efficiency is specified here and measured nowhere yet; it inherits the paper's pre-pilot status.

10.3 The model–evidence frontier

Weak target-specific skill has two distinct causes, and conflating them misdirects effort (the bottleneck argument of Section 5, stated as a design tool). Skill can be low because of an **evidence limitation** — the signal that would resolve the prediction was never captured, so no architecture can recover it — or because of a **model limitation** — useful signal is present in the evidence, but the architecture cannot exploit it. Richer sensing does not fix a model limitation, and a stronger model does not fix an evidence limitation; the two cells on the anti-diagonal of Figure 4 look identical in a single skill number and require opposite investments.

TargetSpace separates the two experimentally, because it varies evidence and architecture on the same sealed instances against the same gates. Hold the model fixed and vary the evidence tier, and the change in gated skill is **observation value** (the evidence-tier ablation). Hold the evidence fixed and vary the architecture class, and the change is **model value** (a column of the grid, Section 7). Vary both and the envelope of best-achievable gated skill traces the **model–evidence frontier**.

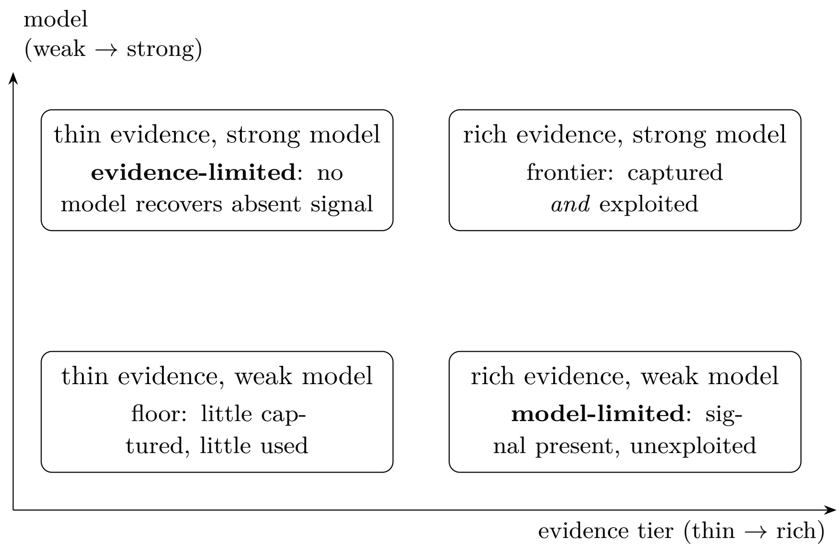

Figure 4: The model–evidence frontier. A single low skill number cannot distinguish an evidence limitation (top-left: the signal was never captured) from a model limitation (bottom-right: the signal is present but unused); the two demand opposite investments. TargetSpace separates them experimentally by varying one axis at a time. The layout is conceptual and carries no data.

Add the disclosed cost and privacy vectors and the frontier becomes a **Pareto** surface over skill, energy, burden, and exposure (Section 10.4). The decomposition is directly useful: it tells a model builder whether a more capable architecture would pay at the current sensing tier, a sensor team whether a new modality would pay before it is built, and a privacy team which streams can be removed with no loss of validated skill; Table 2 maps the full decision set.

10.4 Minimum sufficient observation

The design objective the family points at is **minimum sufficient observation**: a least-burdensome governed evidence configuration that preserves a pre-specified, organizer-set level of gated, target-specific predictive skill. The definition is per disclosed cost dimension — over the full cost vector, whose privacy components are never scalarized, the well-posed object is a Pareto set rather than a unique minimum — and the minimal configuration is identifiable only up to the skill estimate's confidence band. The quantity is task-dependent, and no universal minimum sensor suite exists: for one outcome, calendar metadata may suffice; for another, audio may be necessary; for another, sparse images may add real lift; for some outcomes, no acceptable observation configuration may exist at all, and that finding is itself a benchmark result. Because gated skill is expected to saturate while observation cost keeps rising (Figure 5), the benchmark rewards evidence minimization: when less invasive sensing produces equivalent gated skill, the smaller configuration wins — minimum sufficient governed observation, not maximal surveillance. The incentive is operative rather than aspirational: the reporting contract carries a pre-registered equivalence margin, ties within the margin rank the lower evidence tier first, and every headline result reports the minimal tier whose gated skill falls within the margin of the best (Section 9.5). More invasive capture must earn its place in bits, and when it cannot, it loses the comparison by rule, not by exhortation.

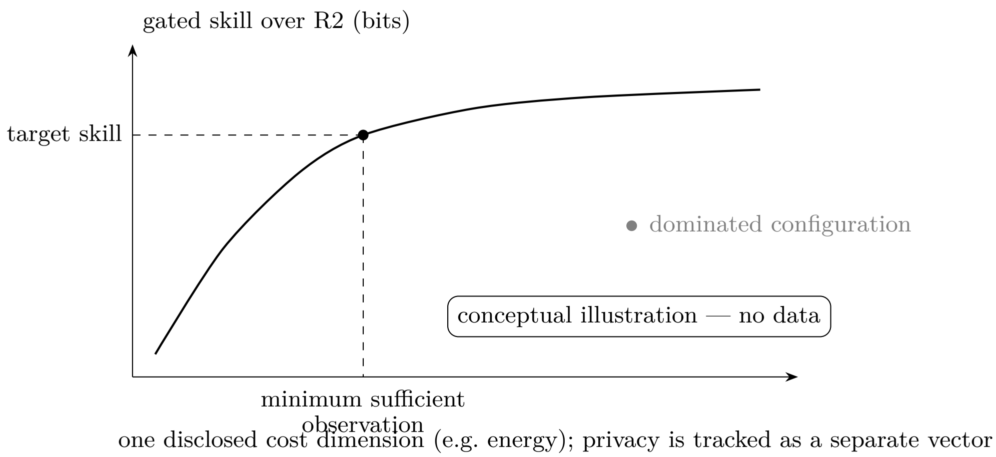

Figure 5: Minimum sufficient observation, conceptually. Gated skill is expected to saturate while observation cost keeps rising; the research objective is the least burdensome governed configuration that preserves a pre-specified skill level, and higher-cost configurations delivering no additional gated skill are dominated. The curve is illustrative and carries no empirical content.

10.5 The observation design space

The agenda is the systematic study of evidence efficiency across the observation design space — modality, cadence, coverage, placement, compute location, and governance — summarized as nine themes in Table 11. Each is answerable with the apparatus already specified: fix the task set, vary the sensing arm, and read the gated skill and its cost. TargetSpace does not assume that richer sensing is better; it asks which evidence configurations remain non-dominated after predictive lift and observation cost are measured together. The optimal observation system is not the one with the highest raw skill but one on the Pareto frontier across skill, energy, coverage, weight, comfort, manual interactions, raw data volume, privacy exposure, latency, cost, exportability, and reliability.

10.6 Transition-centered observation

The benchmark's target is future observable state transitions, so the observation agenda distinguishes routine-state observation, transition-window observation, pre-transition evidence, post-transition evidence, and background evidence that predicts nothing. We state a hypothesis, not a finding: a small fraction of observed time may contain a disproportionate share of transition-relevant predictive information. The pre-registered stratification into routine-continuation and transition cases (Section 9.6) already reports skill separately by stratum; the observation-side experiments extend it — compare random sampling with transition-adjacent sampling, fixed duty cycles with prospectively triggered sensing, and routine-hour with transition-window lift, and quantify value per joule by stratum. One constraint is absolute: a trigger policy is part of the sensing arm, so it must be frozen before evaluation and may not condition on information unavailable at capture time. Adaptive sensing that peeks at outcomes is leakage, not efficiency.

10.7 Reactivity and forgettability

Observation devices can change the behaviour they measure. Sources include visible cameras, recording indicators, manual activation, displays, assistant responses, notifications, social awareness of being recorded, charging rituals, discomfort, and repeated prompts. The naturalistic-observation

Table 11: The observation-science agenda: nine themes, each an open experimental program. Every minimal experiment uses the sealed protocol with pre-registered arms; the primary validity risk names the failure mode the design must control. All themes are open; none has been run.

Theme	Core questions	Minimal experiment / primary risk
A. Sensing sufficiency	Is continuous audio sufficient to beat R2? What do sparse images, mobility, physiology add? Which modalities are redundant, and which matter for deviations rather than routine?	Tier ablation on identical sealed instances / unmatched arms
B. Temporal coverage	How many valid hours per day, and days per study, are required? How does skill degrade with non-wear? Are transition windows worth more than routine hours?	Coverage-stratified re-scoring / coverage confounded with engagement
C. Sampling & duty cycling	Can sparse images replace continuous video? Can event-triggered capture beat fixed schedules? What rate preserves predictive information?	Schedule arms with frozen triggers / outcome leakage through the trigger
D. Placement & form factor	How do chest, wrist, glasses, pendant, earbud, phone, or room placement change the evidence? Which placements minimize reactivity and preserve attribution?	Placement arms, same tier / placement confounded with coverage
E. Compute placement	What must run at the edge? What can be deferred? When does edge summarization destroy evidence needed later?	Raw-vs-derived arms / lossy preprocessing unreported
F. Missingness & uncertainty	How should non-wear, battery failure, and gaps be encoded? Can models distinguish no event from no observation? How does coverage enter calibration?	Explicit missingness manifests / silent imputation
G. Privacy & governance	How much raw evidence must be retained? Can derived features replace raw? How is bystander exposure minimized and derived inference deleted?	Federated runs, derived-only arms / privacy cost hidden from the comparison
H. Longitudinal adaptation	How fast does a model adapt to a new target? How does evidence value decay? How are regime changes distinguished from noise?	History-length arms, walk-forward / drift mistaken for skill
I. Transition-centered observation	Which evidence predicts transitions? Are transition windows rare but disproportionately informative? Can high-value windows be identified prospectively?	Routine/transition strata reported separately (Section 10.6) / post-hoc window selection

literature has managed this for two decades: the Electronically Activated Recorder tradition [67] uses duty-cycled sampling with documented habituation, and its sampling discipline is the default L2 configuration here. **Forgettability** is an experimental property, not a marketing word, and it is defined as minimal burden *subject to* hard disclosure constraints — wearer notice and a bystander-perceptible recording state dominate it. Candidate measures: self-reported awareness of the device, device-interaction counts, removal frequency, non-wear share, and behaviour drift between early and later study periods, read against documented habituation curves [67]. Enrollment consent is not ongoing consent: the protocol includes periodic re-consent, and habituation is a validity resource, never a consent substitute. Visible-versus-concealed comparisons are admissible only where legally and ethically permitted; inconspicuousness never overrides consent, recording law, or bystander rights (Section 13).

10.8 Current hardware as scientific instruments

No shipping product is an optimized observation device, and none needs to be for Layer 1. Table 12 classifies current hardware by experimental role: what each class can test, and the validity risk that most constrains it.

Table 12: Current hardware classified by experimental role. Every class is usable as a Layer-1 instrument for some theme of Table 11; none is claimed to be a good all-day product, and no specific product is endorsed or ranked.

Instrument class	What it can test	Key limit / validity risk
Passive audio wearables	continuous conversational evidence (L2); audio sufficiency; sampling schedules	coverage gaps from battery and social contexts
Meeting recorders	bounded, deterministic outcomes; fast starter studies on commitments and replies	evidence limited to scheduled contexts
Audiovisual glasses / cameras	modality ablation (does vision add lift over audio?); sparse-image substitution	reactivity of visible cameras; short battery
Research sensing platforms	raw data, synchronization, custom acquisition, timestamp quality	not wearable at habitual timescales
Enterprise capture systems	repeated workflow-state transitions on organizational exhaust (L0–L1)	consent and power asymmetry; personal–organizational boundary
Assistive systems	high-stakes calibration and abstention behaviour	heterogeneous capacity; strictest consent architecture
Interaction-first devices	the cost of co-locating observation and intervention (reactivity measurement)	the intervention contaminates the observation

10.9 The ideal observation device

The benchmark's validity conditions imply a design envelope for Layer 2 hardware. We state it as hypotheses to test, not a product specification. The device should run 12–24+ hours between charges, because charging gaps are structured missingness: the hours a device spends on a charger plausibly coincide with exactly the routine breaks the benchmark most needs to observe. It should be passive, comfortable, and forgettable — always worn, no display, minimal interaction — because reactivity is a validity threat (Section 13): a device the wearer performs for measures the performance, not the person. It should carry no interaction surface, because an output channel turns the instrument into an intervention and strains the observe-not-intervene rule that binds scoring windows. It should sense simply and defer inference: an ultra-low-power microcontroller writing compressed evidence to encrypted local storage, uploaded once daily to the wearer's own enclave, keeps the energy budget in capture rather than compute, keeps thermal output and social salience low, and keeps raw data inside the federation boundary of Section 6.3. It should be socially unremarkable in burden, never in disclosure: wearer notice and a bystander-perceptible recording state are hard constraints that dominate forgettability. Venue-based removal — the dinner table, the clinic, the school — is the protocol working as intended, and it enters the record as legitimate structured missingness rather than as a defect in coverage. Beyond these, the instrument must document itself: battery telemetry (enabling EE-energy), a hardware manifest (sensors, firmware, sample rates, duty cycles, trigger logic), explicit missingness logs (protecting calibration analysis from silent gaps), timestamped capture with a declared clock source, physical recording controls, and a visible governance state. These are not product features; they are the difference between a gadget and an instrument. Whether such a device outperforms today's instruments is an empirical question — precisely the one evidence efficiency is defined to answer.

11 Conceptual background and lineage

The idea that prediction is central to modeling is old and well-owned, and this paper claims none of it. In the neural-network tradition, understanding has long been treated as learned internal structure that supports prediction — distributed representations rather than stored symbolic propositions [88]. World-model research makes the same commitment architectural: learned latent models that predict how an environment evolves [15], joint-embedding predictive architectures that predict in representation space [17,18,19,77], and learned models used for planning [20]. Cognitive science supplies converging accounts: predictive processing and the free-energy tradition treat brains as prediction machines [16,21,83], and the memory-prediction framework makes prediction the test of a learned model of the world [84]. Two disciplines bound the claim. Pearl's causal hierarchy places association below intervention and counterfactuals [85], a ranking this paper accepts: prospective prediction from observation sits at the associational rung, strengthened — not replaced — by the counterfactual and intervention modules of the validity stack. And the statistical literature distinguishes predictive from explanatory modeling [74], a distinction the benchmark preserves by scoring only the former. Finally, the risks of optimizing predictive proxies are themselves well documented: Goodhart dynamics [87] and shortcut learning [86] are the standing objections any prediction benchmark must answer, and Section 4 answers them by construction rather than disclaimer.

TargetSpace's novelty is not the claim that prediction matters. Its novelty is making prediction *target-specific, longitudinal, prospective, calibrated, and controlled* — applying the prediction observable to the question of whether a system has learned a specific target, and surrounding it with the validity controls that make the answer meaningful.

11.1 Personal world modeling as a proving ground

World-model research asks whether a system can infer a predictive representation of an evolving environment from partial observation and use it to anticipate what happens next [15,17]. Its most visible instances concentrate on physical scenes, simulated or interactive environments, and robotic control: joint-embedding predictive architectures for images and video [18,19,77], generative interactive-environment models [78,79], learned latent models for planning [20], and generalist robot policies [24]. The analogous problem for a persistent *individual* is comparatively neglected, and it is unusually difficult. Observations are multimodal and incomplete; the dependencies that matter span weeks rather than frames; the target itself changes, so yesterday's model decays; identical outward behaviour can arise from different underlying constraints; population regularity, stable personal routine, and genuine state transitions must be separated rather than conflated; the achievable skill is bounded because much of a person is irreducibly unpredictable [8,9]; and identity and evidence attribution must persist correctly across time, binding each week of evidence to the right target. TargetSpace does not prescribe how any of this is represented internally — a personal world model is one architectural hypothesis among many (Section 2) — it evaluates whether *any* architecture converts partial longitudinal evidence into calibrated, target-specific prospective skill.

The significance is stated carefully: a high score would not establish general intelligence, consciousness, or full psychological understanding (Table 1), and personal world modeling is not claimed to be sufficient for advanced machine intelligence. The defensible claim is structural: building a persistent, calibrated predictive model of one real person from partial longitudinal evidence requires a difficult *conjunction* of capabilities that more capable machine intelligence will likely need more broadly (Table 13). TargetSpace supplies an instrument for measuring progress on that conjunction — simultaneous pressure on all of it, with the failure-mode read-out (Table 7) localizing which element a system lacks. A proving ground, not a proof.

Table 13: The capability conjunction: what target-specific personal prediction demands, and the control that pressures each element — an argument about what the task demands, not a claim that a passing system has mastered any capability in isolation.

Capability	**Why target-specific personal prediction tests it**
Persistent identity	evidence must be bound to the same target across time; the wrong-target permutation gate fails a model whose skill is not about *this* target (Section 7).
Long-horizon memory	individually banal observations become predictive only when integrated across weeks; the longitudinal-history arm rewards integration over recency (Section 9.3).
Temporal abstraction	predictions are scored at multiple horizons, and the shuffled-history control penalizes a model indifferent to *when* evidence arrived.
Multimodal integration	the evidence ladder spans metadata, text, audio, egocentric audiovisual, location, and physiology; the tier ablation scores whether a system fuses them into lift.
Latent-state inference	the scored object is a latent target state read only through observable resolutions; a system must infer state it cannot directly see (Section 6.2).
Adaptation under drift	the target is non-stationary (assumption A3); skill concentrated in transition cases rewards updating over routine replay.
Uncertainty calibration	strictly proper scoring and the calibration gate require honest full-distribution probabilities, not confident point guesses.
Specificity	skill must exceed the target's own routine (R2) and collapse under wrong-target permutation — the two anchors that separate a model of the person from a model of the average.
Selective attention	attention-conditioned lift (Section 8.2) tests whether a system attends to the leading evidence that precedes a transition.
Causal restraint	predictions are scored observationally and causal claims are out of scope; a well-behaved system withholds the unlicensed causal claim (Section 13).
Transfer across tasks	cross-task personal transfer (Section 8.4, hypothesis H6) distinguishes a reusable representation of the person from a narrow per-task rule.

The connection to world-model research is methodological, not reductive, and it is sharpest against the JEPA line of work. JEPA-style systems learn representations that preserve predictable, task-relevant structure without exhaustively reconstructing the sensory surface [18,19,77]. TargetSpace applies the same measurement intuition at the entity level: a personal model need not reproduce the complete record; it must preserve enough target-relevant structure to improve prospective prediction. Table 14 draws the correspondence. Two points prevent it from being read as a rivalry. JEPA is an architectural direction; TargetSpace is an evaluation framework, and a JEPA-style latent predictor is an eligible participant in the grid rather than a competitor (Sections 6.4, 12). And the analogy claims no more than structure: it does not assert that any such system derives explicit laws, that a person is merely a physical scene, or that observation is unbiased — the person remains partially observed, and the benchmark scores only observable future states.

Table 14: Physical and personal world modeling as a structural correspondence, not a reduction: both must preserve predictable, target-relevant structure from partial observation. The person remains partially observed; only observable future states are scored.

Element	Physical world modeling	Personal world modeling
Observed input	partial visual / sensor context of a scene	partial longitudinal personal evidence
Latent object	a representation of the evolving scene or environment	a belief over the evolving target state
Prediction target	a future representation or task outcome	a future observable target-state transition
What is preserved	task-relevant structure, not every pixel	target-relevant structure, not every word or action
Downstream use	interaction, planning, or control	validated assistance or planning, as a separate layer (Section 13)

12 Related Work and Positioning

12.1 What prior work owns, by cluster (adopted, not claimed)

Credibility requires being explicit that TargetSpace invents none of its ingredients: it assembles established tools around one question. We group the neighbours by cluster and keep only the load-bearing comparisons here; the extended discussion and Table 21, itemizing each adopted component and its use, are in Appendix J.

Personal sensing, phenotyping, and lifelogging. The nearest empirical neighbours gather longitudinal individual-level signals from everyday life: experience sampling and ecological momentary assessment [38], personal sensing [39], digital phenotyping [40], egocentric corpora [34,35], and lifelogging with its constructive critique of total capture [58]; the wearable, mobile, ubiquitous, and energy-aware sensing literatures engineered much of the capture layer this agenda depends on; and latent states are demonstrably inferable from passive traces [59]. These are data resources that predict outcomes, whereas TargetSpace makes target-specificity the measured quantity, and the literature on self-report bias [36,44,56,57] and self–other asymmetry [37] motivates treating self-report as auxiliary evidence (Section 5).

User modeling, personalization, and goal recognition. Personalization [25], preference modeling [29], persona and memory benchmarks [28,32], generative agents and digital twins [26,30], a foundation model of cognition [31], and retrieval-augmented or long-context memory [47] predict held-out responses or surface what happened; machine and Bayesian theory of mind [27] and goal or plan recognition [14,33] infer which fixed goal explains behaviour. TargetSpace instead scores whether stored history becomes an improved prediction of the next target-state transition under the permutation gate.

World models and latent predictive learning. World-model agents [15,20], JEPA and its variants [17–19], and predictive coding [21] learn to predict in a representation space; TargetSpace evaluates such systems as an architecture class rather than replacing them, scoring the externally resolved transition, not the representation.

Forecasting, scoring, calibration, sealing, and contamination. Proper scoring [10–12], calibration as a measured axis [6], sealed prospective event forecasting [13], and contamination-resistant private evaluation [3,4,7] are adopted directly; they score external public events or fixed tasks, not the target-state transitions of a tracked instance under an own-routine baseline and a permutation gate.

Ethics of ambient capture and inferred data. A distinct literature governs the risks: bystander and recording ethics [61], the privacy status of inferred rather than disclosed data [60], manipulation via inferred state [62], workplace surveillance [64], and behavioural prediction as a traded product [63]. These bear on the governance boundary of Section 13, where the derived inference object, not only the raw signal, is the privacy boundary.

12.2 The conjunction

Each ingredient in Table 21 is occupied somewhere; the conjunction is not. We state the supported claim narrowly: **to our knowledge no existing benchmark scores prospective, calibrated, proper-scored predictions of a tracked target system's latent target-state transitions under a strong instance-specific own-routine baseline (R2) and an instance-permutation specificity gate, with an evidence-tier ablation, across architecture classes, linking the choice of evidence configuration to the gated predictive skill it produces.** We do *not* claim to be first at architecture-neutral evaluation [22], cross-architecture comparison [23], calibrated sealed prediction [13], evidence ablation, or resolution-timing [14]. The contribution is their assembly around one measurable question.

A structured comparison of neighbouring benchmarks against the five dimensions (Appendix J, Table 22) shows the pattern: prior work holds important *subsets* — sealed proper-scored prediction (ForecastBench), a persistent individual (PersonaMem, Generative Agents, KnowMe-Bench), latent-goal inference (EgoToM, ToMnet, goal/plan recognition), novel-task adaptation (ARC-AGI) — but to our knowledge none combines all five in one prospective apparatus. The labels are cautious design judgments, not measured scores, and the TargetSpace row reflects its proposed design, not completed validation. The contrast with ARC-AGI [7] is representative: ARC asks whether a system can acquire the skill to solve a novel task whose frame and success condition are given; TargetSpace asks the earlier question of whether a system can infer the relevant frame — the latent target and its likely transition — when nothing is framed for it. Both are needed; neither subsumes the other.

12.3 Positioning

Two caveats specific to this positioning (the general limitations are Section 13): TargetSpace is *not* a substitute for physical-reasoning, robotics, or video-realism benchmarks, and a strong score implies nothing about those capabilities; and models may overfit to retrieval-like cues — surfacing *what* happened — rather than learning target dynamics, which is why higher recall or longer context does not by itself move prospective skill over R1/R2 under the permutation gate (Section H.1). Predictability itself is bounded [8,9], so no single headline number can represent the construct; skill is reported per instance, against R1/R2, under the calibration and permutation gates.

13 Limitations, Ethics, and Governance

Behavioural prediction is not moral authority. We keep benchmark validity separate from deployment legitimacy. A high TargetSpace score certifies calibrated, prospective, target-specific predictions about a consenting individual's future observable states under sealed conditions; it grants no access to consciousness, intent, or inner life (never scored), and it is not permission to act on, steer, manipulate, or surveil the person, nor a licence for employment [64], insurance, or coercive use [62,63]. A calibrated behaviour predictor is dual-use, and the observe-not-intervene rule binds the benchmark, not downstream deployers; a containment/refusal axis alongside the skill axis is required future work. A positive result establishes predictive skill, not understanding in a rich sense, and skill measured under passive observation is predictive, not causal, since interventional identification would be needed.

Principal limitations. The most consequential: latent targets are evaluated only through observable proxies; the federated design is holder-reproducible rather than publicly auditable; small-cohort evaluation, selection bias, and overfitting to one idiosyncratic target limit external validity, since the permutation gate guards cross-target leakage but not unrepresentative sampling; capture can be reactive, so habituation windows and reactivity checks are required and induced deviations are not read as transitions; targets drift, so R2 is re-fit walk-forward; predictability is bounded [8,9], so reporting is per-instance; and the privacy and governance safeguards are design commitments, not implemented features (no pilot has run, no review obtained). We specify informed revocable consent, federation so raw data never leaves the target's control, aggregate-only reporting, institutional review before any deployment, and respect for recording law and bystander consent [61]; a benchmark that rewarded changing the target would measure influence, not understanding. The observation agenda adds governance surface: sensor minimization is an objective rather than an afterthought (Section 10.4), non-wear must remain a costless choice rather than a penalized deviation, energy and environmental cost are disclosed rather than externalized, and subgroup calibration differences are treated as validity concerns before any deployment question arises. The full register (L1–L16), unresolved issues, and a minimal safe pilot configuration are in Appendices F and E; no public raw first-person dataset is released, and the synthetic harness and protocol are the released artifacts.

13.1 The inference object, not the raw signal, is the privacy boundary

Privacy controls focused only on raw audio or video are incomplete. Even if recordings are deleted, an ambient target-state system (Section 6) retains transcripts, embeddings, entity maps, commitments, routines, and inferred goals, derived objects often more searchable and actionable than the recording itself (Table 27); their distinct legal and privacy status is the subject of a growing derived-inference literature [60]. Deleting the recording does not delete the target-state risk if these layers persist, and the same layering concentrates bystander, purpose-drift, and power-asymmetry risks. The boundary applies to the benchmark's own exports: per-instance sealed predictions and resolved labels are themselves derived inference objects about an identified participant. By default they remain in the holder's enclave; what crosses is aggregate reporting above a pre-registered minimum cohort size, and the five-person feasibility cohort is below any releasable threshold — its per-instance outputs are holder-internal by design [60]. Local processing and open-source implementations are valuable mitigations [49,54], not complete solutions. These systems also have clearly beneficial uses (accessibility, memory support, care coordination, safety), which is why a shared measurement and governance vocabulary is needed rather than alarm; the industry cluster is in Appendix M.

14 What success and failure would mean

Because the benchmark is falsifiable, it is worth stating in advance what either outcome would teach, so that a null result is read as evidence rather than as the instrument's failure.

If the principal hypotheses hold — skill beyond R1 and an admitted R2, surviving calibration and permutation, concentrated in transition cases, and rising with observation tier (H2–H5, Section L.2) — it would support several claims that are today asserted rather than measured: that there is target-specific information about a person beyond population base rates and personal routine; that longitudinal evidence carries extractable *dynamic* signal, not only a static profile; that observation modalities can be *ranked* by the predictive value they add; that a claimed personal model can be *audited* against sealed outcomes; that sensing can be designed to minimize evidence for a target skill level; and that reusable personal representations are within reach (H6). None of this would establish understanding, consciousness, or permission to act (Table 1); it would establish that a measurable capability exists and can be improved against.

A well-designed null is equally informative. Persistent failure to beat R2, or skill that does not survive the gates, would be consistent with several distinct diagnoses, and the control battery (Section 9.3) is built to tell them apart: that current personal-AI systems are largely performing retrieval and routine replay rather than modeling the person; that available sensors do not capture the state the outcomes depend on (an evidence limitation, Section 10.3); that current architectures cannot use the evidence that is present (a model limitation); that the chosen outcomes are too intrinsically unpredictable to score [8,9]; that observation coverage is insufficient; or that the target-state definitions need revision. Each is a finding. The uninformative outcome is not a null — it is an unaudited claim, which is the state personal AI is in now; a clean null would already be progress, because it would replace a marketing assertion with a measured bound.

15 Conclusion

Personal AI is drifting toward evaluations of recall, memory, and stated-preference satisfaction, which test how well a system re-serves what a person already externalized. TargetSpace proposes a different bar: sealed, calibrated predictions of a specific target's future observable state transitions. R1 removes population base rates; R2 removes routine replay; the permutation gate tests whether the prediction is about this target rather than the average; and the evidence-tier ablation measures what passive observation adds over self-report. The R2 baseline and the permutation gate are the anchors, and the rest is adopted and cited. This is a pre-pilot protocol with a synthetic harness and no human-subject results, and a high score certifies calibrated predictive skill about future observable states, not understanding in any richer sense, inner life, or permission to act. The record is not the target; TargetSpace asks whether a system can predict the target's next transition rather than merely summarize its traces. TargetSpace is therefore also a research-direction claim: progress in personal AI should not be measured only by richer memory or more fluent personalization, but by prospective, calibrated skill at predicting target-specific transitions under controlled baselines and specificity tests. A field that measures memory will build better memory; a field that measures calibrated target-specific prediction will build systems that model the people they serve. TargetSpace forces progress in target modeling, not merely model scaling, memory storage, or data capture. The benchmark is the centerpiece; the convergence map (Section 2), the three-layer sequence, and evidence efficiency (Section 10) exist to make its question the one the field optimizes. Predict the target, not the average, from what can be observed rather than what is professed: the target is not a profile or a self-description, but an observed trajectory in motion. Generic AI evaluations ask whether a model knows the world; TargetSpace asks whether a model has learned *this* target. It does not define all of understanding. It makes one indispensable part of understanding measurable: whether a system has learned a calibrated, target-specific model with predictive force over time.

Declarations

Author contributions. Yuri Andrade Sylvester conceived the TargetSpace framework, developed the benchmark specification and proposed evaluation protocol, conducted the literature synthesis, and wrote and revised the manuscript.

Funding. This research received no external funding and was conducted using the author's personal resources.
Competing interests. The author declares no competing interests.
Ethics statement. No human-participant data were collected for this paper. Dataset collection has not begun, and no ethics approval has been sought or obtained for the proposed future study. Any future participant research implementing this protocol will require appropriate ethics review and informed consent before recruitment or data collection.
Data availability. No participant dataset was generated or analyzed for this paper.
Code and harness availability. The Version 1.0 benchmark protocol and the minimal synthetic reference harness are available through the project website (`https://targetspace.org`) and the public repository (`https://github.com/yurisyl/targetspace-bench`), which also provides the machine-readable JSON Schemas for participants, evidence manifests, predictions, outcomes, submissions, and leaderboards together with worked example records; the harness accompanies this preprint as an ancillary file (Appendix B). The synthetic harness verifies formatting, metric computation, baseline execution, calibration checks, permutation controls, and evidence-ablation behaviour; it uses no participant data, supports no empirical claims, and is not a substitute for participant-derived longitudinal data. Real-world submissions must follow the schema, privacy, consent, and evaluation requirements of the protocol. No participant dataset is released with Version 1.0; pilot outputs and validated datasets are future work (Section 13).
Project website. The TargetSpace project website is available at `https://targetspace.org` and provides the public protocol, benchmark harness materials, documentation, and submission information for the Version 1.0 release. This manuscript is a non-peer-reviewed preprint prepared for deposit on arXiv, a preprint repository rather than a journal; an arXiv identifier is forthcoming.
Correspondence. Yuri Andrade Sylvester, `yurisyl@gmail.com`.

References

A. Benchmarks and evaluation methodology

[1] J. Deng et al. ImageNet: A Large-Scale Hierarchical Image Database. CVPR, 2009.
[2] C. E. Jimenez et al. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? ICLR, 2024. arXiv:2310.06770.
[3] C. White et al. LiveBench: A Contamination-Limited LLM Benchmark. 2024. arXiv:2406.19314.
[4] L. Zhang et al. SWE-bench Goes Live! (the SWE-bench-Live contamination-resistant benchmark). 2025. arXiv:2505.23419.
[5] M. Chen et al. Evaluating Large Language Models Trained on Code (HumanEval). 2021. arXiv:2107.03374.
[6] P. Liang et al. Holistic Evaluation of Language Models (HELM). TMLR, 2023. arXiv:2211.09110.
[7] F. Chollet, M. Knoop, G. Kamradt, B. Landers, H. Pinkard. ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems. 2025. arXiv:2505.11831.

B. Forecasting, calibration, scoring, goal recognition

[8] A. Bellot, J. Richens, T. Everitt. The Limits of Predicting Agents from Behaviour. ICML, 2025. arXiv:2506.02923.
[9] C. Song, Z. Qu, N. Blumm, A.-L. Barabasi. Limits of Predictability in Human Mobility. Science, 327(5968):1018–1021, 2010.
[10] I. J. Good. Rational Decisions. J. Royal Statistical Society B, 14(1), 1952.

[11] T. Gneiting, A. E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. JASA, 102(477), 2007.
[12] G. W. Brier. Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review, 78(1), 1950.
[13] E. Karger et al. ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities. ICLR, 2025. arXiv:2409.19839.
[14] S. Keren, A. Gal, E. Karpas. Goal Recognition Design. ICAPS, 2014.
[74] G. Shmueli. To Explain or to Predict? Statistical Science, 25(3):289–310, 2010.
[75] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger. On Calibration of Modern Neural Networks. ICML, 2017. arXiv:1706.04599.
[81] Q. Yang, S. Mahns, S. Li, A. Gu, J. Wu, H. Xu. LLM-as-a-Prophet: Understanding Predictive Intelligence with Prophet Arena. 2025. arXiv:2510.17638.
[82] L. Nel. KalshiBench: Evaluating Epistemic Calibration of Large Language Models via Prediction Markets. 2025. arXiv:2512.16030.
[85] J. Pearl. The Seven Tools of Causal Inference, with Reflections on Machine Learning. Communications of the ACM, 62(3):54–60, 2019.
[86] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F. A. Wichmann. Shortcut Learning in Deep Neural Networks. Nature Machine Intelligence, 2:665–673, 2020. arXiv:2004.07780.
[87] D. Manheim, S. Garrabrant. Categorizing Variants of Goodhart's Law. 2018. arXiv:1803.04585.

C. World models, JEPA, predictive learning

[15] D. Ha, J. Schmidhuber. World Models. NeurIPS, 2018. arXiv:1803.10122.
[16] K. Friston. The Free-Energy Principle: A Unified Brain Theory? Nat. Rev. Neuroscience, 11, 2010.
[17] Y. LeCun. A Path Towards Autonomous Machine Intelligence. v0.9.2, 2022. OpenReview BZ5a1r-kVsf. (Position paper; not arXiv.)
[18] M. Assran et al. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture (I-JEPA). CVPR, 2023. arXiv:2301.08243.
[19] M. Assran et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. 2025. arXiv:2506.09985.
[20] J. Schrittwieser et al. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model (MuZero). Nature 588, 2020. arXiv:1911.08265.
[21] R. P. N. Rao, D. H. Ballard. Predictive Coding in the Visual Cortex. Nature Neuroscience, 2(1), 1999.
[22] A. Warrier et al. Benchmarking World-Model Learning with Environment-Level Queries (AutumnBench / WorldTest). 2025. arXiv:2510.19788.
[23] D. Chen et al. WorldPrediction: A Benchmark for High-level World Modeling and Long-horizon Procedural Planning. 2025. arXiv:2506.04363.
[24] Physical Intelligence. $\pi_{0.7}$: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities. 2026. arXiv:2604.15483.
[77] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, N. Ballas. Revisiting Feature Prediction for Learning Visual Representations from Video (V-JEPA). 2024. arXiv:2404.08471.
[78] J. Bruce et al. Genie: Generative Interactive Environments. ICML, 2024. arXiv:2402.15391.
[79] Google DeepMind. Genie 2: A Large-Scale Foundation World Model. Research announcement, December 2024. Accessed July 2026.

[83] A. Clark. Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. Behavioral and Brain Sciences, 36(3):181–204, 2013.
[84] J. Hawkins, S. Blakeslee. On Intelligence. Times Books, New York, 2004.
[88] G. E. Hinton. Learning Multiple Layers of Representation. Trends in Cognitive Sciences, 11(10):428–434, 2007.

D. Person modeling, theory of mind, personalization, observation

[25] A. Salemi et al. LaMP: When Large Language Models Meet Personalization. SIGIR, 2024. arXiv:2304.11406.
[26] J. S. Park et al. Generative Agents: Interactive Simulacra of Human Behavior. UIST, 2023. arXiv:2304.03442.
[27] N. C. Rabinowitz et al. Machine Theory of Mind (ToMnet). ICML, 2018. arXiv:1802.07740.
[28] B. Jiang et al. Know Me, Respond to Me: Benchmarking LLMs for Dynamic User Profiling (PersonaMem). COLM, 2025. arXiv:2504.14225.
[29] S. Zhao et al. Do LLMs Recognize Your Preferences? (PrefEval). ICLR, 2025. arXiv:2502.09597.
[30] R. Li et al. How Far are LLMs from Being Our Digital Twins? (BehaviorChain). Findings of ACL, 2025. arXiv:2502.14642.
[31] M. Binz, E. Schulz et al. A Foundation Model to Predict and Capture Human Cognition (Centaur). Nature, 2025. arXiv:2410.20268.
[32] T. Wu et al. KnowMe-Bench: Benchmarking Person Understanding for Lifelong Digital Companions. 2026. arXiv:2601.04745.
[33] Y. Li et al. EgoToM: Benchmarking Theory of Mind Reasoning from Egocentric Videos. Meta, 2025. arXiv:2503.22152.
[34] K. Grauman et al. Ego4D: Around the World in 3,000 Hours of Egocentric Video. CVPR, 2022. arXiv:2110.07058.
[35] K. Grauman et al. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. CVPR, 2024. arXiv:2311.18259.
[36] R. E. Nisbett, T. D. Wilson. Telling More Than We Can Know: Verbal Reports on Mental Processes. Psychological Review, 84(3):231–259, 1977.
[37] S. Vazire. Who Knows What About a Person? The Self–Other Knowledge Asymmetry (SOKA) Model. J. Personality and Social Psychology, 98(2):281–300, 2010.
[38] S. Shiffman, A. A. Stone, M. R. Hufford. Ecological Momentary Assessment. Annual Review of Clinical Psychology, 4:1–32, 2008.
[39] D. C. Mohr, M. Zhang, S. M. Schueller. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. Annual Review of Clinical Psychology, 13:23–47, 2017.
[40] J.-P. Onnela, S. L. Rauch. Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. Neuropsychopharmacology, 41(7):1691–1696, 2016.
[71] A. Rosenblueth, N. Wiener, J. Bigelow. Behavior, Purpose and Teleology. Philosophy of Science, 10(1):18–24, 1943.
[72] G. A. Miller, E. Galanter, K. H. Pribram. Plans and the Structure of Behavior. Holt, 1960.
[73] D. C. Dennett. The Intentional Stance. MIT Press, 1987.
[76] J. D. Mayer. Personal Intelligence: The Power of Personality and How It Shapes Our Lives. Farrar, Straus and Giroux, 2014.

E. Cross-scale target-state systems

[41] M. Levin. Technological Approach to Mind Everywhere (TAME). Frontiers in Systems Neuroscience, 16:768201, 2022.

F. Attention and goal dynamics

[42] E. I. Knudsen. Fundamental Components of Attention. Annual Review of Neuroscience, 30:57–78, 2007.

[43] M. Corbetta, G. L. Shulman. Control of Goal-Directed and Stimulus-Driven Attention in the Brain. Nature Reviews Neuroscience, 3(3):201–215, 2002.

G. Memory, belief states, and goal inference

[44] M. A. Conway, C. W. Pleydell-Pearce. The Construction of Autobiographical Memories in the Self-Memory System. Psychological Review, 107(2):261–288, 2000.

[45] P. Sheeran, T. L. Webb. The Intention–Behavior Gap. Social and Personality Psychology Compass, 10(9):503–518, 2016.

[46] L. P. Kaelbling, M. L. Littman, A. R. Cassandra. Planning and Acting in Partially Observable Stochastic Domains. Artificial Intelligence, 101(1–2):99–134, 1998.

[47] P. Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS, 2020. arXiv:2005.11401.

[48] I. T. Jolliffe, D. B. Stephenson (eds.). Forecast Verification: A Practitioner's Guide in Atmospheric Science. 2nd ed., Wiley, 2012.

[49] Based Hardware. Omi: an Open-Source AI Wearable. Product manifesto and documentation, 2025. Accessed July 2026.

[50] Bee. An Ambient Wearable AI Assistant that Turns Everyday Conversations into Summaries, Reminders, and To-Dos. Product documentation, 2025; reported acquisition by Amazon announced July 2025. Accessed July 2026.

[51] PLAUD. Note and NotePin: Wearable AI Voice Recorders and Note-Takers. Product documentation, 2024. Accessed July 2026.

[52] Limitless. Pendant: a Wearable AI for Personalized Memory over Everyday Conversations. Product documentation, 2024; reported acquisition by Meta announced December 2025. Accessed July 2026.

[53] Meta. Ray-Ban Meta and Meta Ray-Ban Display AI Glasses. Product documentation, 2023–2025. Accessed July 2026.

[54] Brilliant Labs. Halo: Open-Source AI Glasses with the Noa Agent and Cross-Session Narrative Memory. Product documentation, 2025. Accessed July 2026.

[55] Snap. Spectacles and Specs: Standalone See-Through AR Glasses. Product documentation, 2024–2026. Accessed July 2026.

H. Self-report bias, lifelogging, ambient capture, and inferred data

[56] D. P. Crowne, D. Marlowe. A New Scale of Social Desirability Independent of Psychopathology. Journal of Consulting Psychology, 24(4):349–354, 1960.

[57] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, N. P. Podsakoff. Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. Journal of Applied Psychology, 88(5):879–903, 2003.

[58] A. J. Sellen, S. Whittaker. Beyond Total Capture: A Constructive Critique of Lifelogging. Communications of the ACM, 53(5):70–77, 2010.
[59] M. Kosinski, D. Stillwell, T. Graepel. Private Traits and Attributes are Predictable from Digital Records of Human Behavior. Proceedings of the National Academy of Sciences, 110(15):5802–5805, 2013.
[60] S. Wachter, B. Mittelstadt. A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. Columbia Business Law Review, 2019(2):494–620, 2019.
[61] T. Denning, Z. Dehlawi, T. Kohno. In Situ with Bystanders of Augmented Reality Glasses: Perspectives on Recording and Privacy-Mediating Technologies. CHI, 2014.
[62] D. Susser, B. Roessler, H. Nissenbaum. Online Manipulation: Hidden Influences in a Digital World. Georgetown Law Technology Review, 4(1):1–45, 2019.
[63] S. Zuboff. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs, New York, 2019.
[64] I. Ajunwa, K. Crawford, J. Schultz. Limitless Worker Surveillance. California Law Review, 105(3):735–776, 2017.
[65] A. Bulling, U. Blanke, B. Schiele. A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors. ACM Computing Surveys, 46(3):1–33, 2014.
[66] N. Eagle, A. Pentland. Reality Mining: Sensing Complex Social Systems. Personal and Ubiquitous Computing, 10(4):255–268, 2006.
[80] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, A. T. Campbell. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones. UbiComp, 2014. doi:10.1145/2632048.2632054.
[67] M. R. Mehl. The Electronically Activated Recorder (EAR): A Method for the Naturalistic Observation of Daily Social Behavior. Current Directions in Psychological Science, 26(2):184–190, 2017.
[68] R. Hoyle, R. Templeman, S. Armes, D. Anthony, D. Crandall, A. Kapadia. Privacy Behaviors of Lifeloggers Using Wearable Cameras. UbiComp, 2014.
[69] P. Kelly et al. An Ethical Framework for Automated, Wearable Cameras in Health Behavior Research. American Journal of Preventive Medicine, 44(3):314–319, 2013.
[70] H. Nissenbaum. Privacy as Contextual Integrity. Washington Law Review, 79(1):119–158, 2004.

A Glossary of core terms

Definitions are deliberately compact; the body is authoritative where they differ.

Target state. A latent configuration a target behaves as if acting to reach, maintain, resume, or abandon; evaluated only through observable consequences, never scored directly. **Stated goal.** What a target declares it wants — evidence, not a scored quantity. **Enacted priority.** What a target's allocation of a scarce resource (attention, time) reveals, independent of what it states. **Inferred goal.** What a model concludes a target is pursuing; evidence-bearing latent structure, scored only via a mapped outcome. **Implicit target.** A latent orientation inferred from behaviour, repetition, avoidance, or attention that the target has not declared and may not introspect. **Score target.** A pre-registered discrete future state or transition with an externally observable resolution — the only object that earns or loses credit; a candidate lacking such a rule is **evidence**, not a score target. **Inclusion rule (hard).** A target state is included in TargetSpace only if it has a pre-registered observable resolution rule; otherwise it is evidence, not a scored state (Section 6.2, Table 4). **Sealed outcome (outcome).** The realized value of a score target at the resolution time, fixed by a deterministic rule and hashed before it exists; scoring compares the sealed prediction against it. **Prediction horizon.** The interval $h = r - t$ from the sealed prediction time t to the resolution time r; predictions are specified and scored per horizon. **Transition window.** The span over which a target-state transition is deemed to occur for resolution purposes, distinct from the prediction horizon. **Belief state.** The observer/model's distribution over a target's latent variables, updated as evidence arrives; a sufficient statistic of the history that the benchmark motivates but never scores. **Transition.** A change in which target state a system is acting to reach: emergence, stabilization, competition, displacement, abandonment, resumption, or opportunity-induced creation. **Population-prior baseline (R1).** A reference that predicts from population base rates with no target-specific information; the entry condition a model must beat. **Own-routine baseline (R2).** A pre-registered, walk-forward model of the target's own recent routine; skill over R2 is the headline target-specific signal. **Evidence tier.** A cumulative band of the evidence ladder (Section L.2) describing which streams a system may use, from low-content metadata to specialized sensors. **Evidence ablation.** Measuring skill over R2 as a function of evidence tier, to test which streams add target-specific information. **Permutation specificity gate.** A test that scores a system's predictions for one target against another target's outcomes; genuine target-specific skill should collapse. **Prospective sealing.** Timestamping and hashing (SHA-256) a prediction before its outcome exists, preventing hindsight and leakage. **Walk-forward evaluation.** Prequential scoring using only evidence timestamped at or before t; random cross-validation is prohibited. **Calibration gate.** A pass/warn/fail check that stated confidence tracks empirical frequency (proper scoring plus expected-calibration-error bands). **Target-specific skill.** Calibrated predictive skill, in bits, that exceeds both R1 and R2 and collapses under permutation. **Epistemic resolution.** The falling of the observer's uncertainty as evidence accumulates (posterior concentration), distinct from participatory formation in the target. **Participatory dynamics.** Observable feedback, attention, action, and constraint signals through which a target's own process may change which futures remain reachable; recorded as auxiliary diagnostic variables (Appendix G), never part of primary scoring.

B Minimal synthetic harness and submission specification

Purpose and scope. The minimal synthetic harness is a public, runnable reference implementation for validating instance and forecast schemas, exercising proper-score and Skill computation, checking the R1 and R2 baselines, running the permutation control, running calibration checks, running evidence ablations, and producing a submission report. It is a smoke test plus reference path: it is *not* the empirical validation dataset, and it is *not* evidence that the benchmark has been experimentally validated. It uses no human data and supports no empirical claims. The harness and this specification are distributed through the project website (`https://targetspace.org`) and the public repository (`https://github.com/yurisyl/targetspace-bench`), and accompany the preprint as ancillary files.

What it does not demonstrate. No human data; no real personal intelligence; no evidence that passive observation helps; no cross-domain validation; no safety validation.

Required input schema (instances and evidence). Per forecast instance: `instance_id`; anonymized `target_id` (`subject_id`); `forecast_time` and `resolution_time` (or an explicit `horizon`); `question_type`; `answer_space`; the deterministic `resolution_rule`; and a `consent_status` / eligibility flag. Per evidence record: `segment_id`; time window; `modality` label (evidence tier L0–L6); optional text or transcript summary; and provenance pointers. Out-

Table 15: Protocol elements (referenced from Section 6.2): each a concrete definition with its pre-registered option(s). Design choices are fixed before a prediction is sealed and disclosed with the result.

Element	Definition	Pre-registered option(s)
Target	the tracked instance i	person / agent / system / organization / process
Target state	latent configuration z_t the target acts to reach or maintain	scored only via observable consequences (A2)
Prediction	a distribution over the discrete answer space A at time t	proper-scored; nonzero probability floor
Observation window	evidence $E_{\leq t}$ available up to predict time t	zero- / short- / longitudinal-history arms
Prediction horizon	$h = r - t$, from t to resolution time r	short / medium / long; scored per horizon
Label / outcome	the resolved value at r	set by a deterministic resolution rule
Ground truth	how the outcome is established	observed action, sensor threshold, or pre-registered rule; inter-rater reliability where contestable
Evaluation metric	how skill is scored	log score (bits), Brier; Skill over R1/R2
Baseline	reference the system must beat	R1 population prior; R2 own-routine
Uncertainty reporting	calibration and interval reporting	calibration gate; day-blocked / person-clustered intervals
Leakage control	preventing use of future information	sealing (SHA-256) + strict walk-forward; random CV prohibited

come labels are withheld from the system under test and stored separately (`instance_id`, resolved `outcome`, resolution provenance). Target-state or transition markers used to construct instances follow the taxonomy of Appendix I. In the shipped TS-Personal JSON Schemas (the released TargetSpace schema set) these generic fields instantiate concretely: the target is `participant_id`, the instance is `task_id`, the question type is `task_type`, the resolved label is `observed_answer`, and evidence modality/window are `modalities` with `evidence_cutoff_time`; the abstract vocabulary here is the domain-general layer over that concrete personal-track schema.

Required output schema (per system). Per forecast: `forecast_id`, `system_id`, `instance_id`, `forecast_time`, a probability for every element of `answer_space` (ranked candidates permitted for ordered spaces); a payload `sha256` seal recorded in the run/sealing manifest keyed by `forecast_id` (a hash of the payload is stored alongside the record it seals, not inside it); and optional `evidence_refs`, carried in the prediction `metadata`, identifying the segments supporting the prediction. Per run: a metrics file (JSON or CSV) with the scores below; a run manifest (data slice, config, seeds, timestamps); a model card or system description including the output adapter, retrieval/memory access, and training cutoff; an ablation report per evidence tier; and the baseline comparison against R1 and R2.

Baselines. The harness ships five reference conditions: **R0**, a uniform-random / prevalence floor (sanity only); **R1**, the population-prior baseline fit leave-one-out without the scored target's history; **R2**, the target's own-routine baseline (default recipe in Appendix D); the **target-specific system** under test; and the **permutation null**, the system's predictions scored against matched wrong-target outcomes. Evidence-ablation variants re-run the system per evidence tier. R2 is the decisive reference: a claimed TargetSpace improvement must beat R2 — not merely R0 or generic replay — and its skill must collapse under the permutation null.

Metrics. The harness reports, per horizon and per evidence tier: log score (bits; primary) and Skill over R1 and over R2; Brier score; calibration (top-label ECE bands with pass/warn/fail, plus intercept/slope where sample size permits); the permutation-control delta (true-pairing Skill minus wrong-pairing Skill); and the evidence-ablation delta per tier. For binary or ranked answer spaces, AUROC/AUPRC and top-k accuracy may be reported as secondary diagnostics; no composite score is used, and lead-time (horizon-specific) performance is reported rather than pooled (Appendix K).

Command-line flow. The single-file demonstration is runnable today as `python targetspace_synthetic_demo.py` and executes the full flow (generation, walk-forward evaluation, R1/R2, scoring, calibration, permutation, ablation) in one run. The stepwise commands below are the *reference CLI specification* — the intended command structure for the full submission harness; they are *not part of the arXiv ancillary bundle*, which ships only the single-file harness (a fuller reference implementation is maintained in the public repository):

- `python generate_synthetic_targets.py -config example_config.yaml` (or: validate your own data against the input schema)
- `python run_walk_forward_eval.py -config example_config.yaml -baselines r0,r1,r2`
- `python scoring.py -predictions runs/system/forecasts.jsonl -truth data/outcomes.jsonl`
- `python permutation_test.py -predictions runs/system/forecasts.jsonl -matching matched`
- `python run_walk_forward_eval.py -config example_config.yaml -ablate evidence_tier`

Ancillary files (shipped with Version 1.0). `README.md` (run instructions and scope); `targetspace_synthetic_demo.py` (the runnable single-file harness; Python 3.8+, standard library only, deterministic); `example_output.md` (a captured reference run); `requirements.txt` (no third-party dependencies). The stepwise standalone-script invocation named above (`generate_synthetic_targets.py`, `run_walk_forward_eval.py`, `permutation_test.py`, `example_config.yaml`) is a reference *specification* and is not part of the arXiv ancillary bundle, which ships only the single-file harness. The corresponding capabilities — proper scoring, the R1/R2 baselines, and the permutation and calibration checks — are nonetheless implemented in the public repository as an installable package and command-line tool, on synthetic data only and supporting no empirical claim.

Submission readiness. A team with compliant longitudinal passive-observation data should be able to use this specification to format a submission, run the reference checks, compare against R1/R2, and generate a benchmark report. The synthetic harness verifies the mechanics; it does not itself constitute a benchmark result.

Table 16: Illustrative synthetic leaderboard row produced by the demo script. Skill is measured in bits over the named reference baseline.

system_id	tier	vs_R1	vs_R2	brier	cal. warn	perm. loss	n_tgt	n_fc
toy-L2	L2	+0.049	+0.036	0.183	none	collapses	20	800

Example leaderboard row. Values are illustrative synthetic outputs of the demo script; they are not empirical results.

The columns report, in order, the system identifier, evidence tier, skill in bits over R1, skill in bits over R2, Brier score, calibration-gate warning status, the permutation specificity outcome (skill collapses under permutation, as required), and the target and prediction counts.

C Running TargetSpace on a product: implementation guide

This appendix is an operational recipe for a product team — a note-taking or memory app, a wearable audio recorder, or multimodal glasses — to run a TargetSpace evaluation on its own users' data without guessing. It reuses the released schemas (Appendix B) and adds no new record types. Everything below is a protocol usage guide, not an empirical claim; a run over real users must satisfy the consent, privacy, and governance requirements of Section 13 and Appendix E.

Minimum-viable benchmark path

The smallest defensible run is seven steps; each maps to one released TargetSpace schema.

1. **Choose target instances.** Select consenting users and a sealed evaluation window. Register each as a target with `participant.schema.json` (`participant_id`, `timezone`, `routine_profile`, `allowed_tasks`); for real users the synthetic fields are replaced by de-identified profile metadata.
2. **Choose a task family.** Pick one or more rows from the quickstart pack below. Each fixes a `task_type`, an `answer_space`, and a deterministic `resolution_rule`.
3. **Freeze evidence tiers.** Declare which evidence tiers (L0–L6, Section L.2) each condition may use in `evidence_manifest.schema.json` (`evidence_tier`, `modalities`, `evidence_start_time`, `evidence_end_time`, `hash_manifest`). One manifest per tier condition drives the evidence ablation.
4. **Generate sealed predictions.** Before any outcome is observed, emit one `forecast.schema.json` record per instance: a probability over every element of `answer_space`, an `evidence_cutoff_time` $\leq T$, and a hash sealing the payload. Store forecasts write-once.
5. **Resolve outcomes deterministically.** At resolution time, apply the task's resolution rule to later observable evidence and write `outcome.schema.json` (`observed_answer`, `resolver`, `resolution_method`). Outcomes are withheld from the system under test until sealing.
6. **Score against R1 and R2.** Compute log score (bits, primary) and Brier per instance; report Skill in bits over the population prior R1 and over the own-routine baseline R2 (Appendix D), plus top-label calibration (ECE band). A claimed improvement must beat R2, not merely R1 or replay.

7. **Run the permutation control and report the standard row.** Re-score each system against matched wrong-target outcomes; target-specific skill must collapse. Emit one `leaderboard.schema.json` row (and a `submission.schema.json` record) with the columns of Table 16: `system_id`, `tier`, `vs_R1`, `vs_R2`, `brier`, calibration status, permutation outcome, `n_tgt`, `n_fc`.

Quickstart task pack

Six product-relevant task families, each expressible in `forecast.schema.json` with the answer space and deterministic resolution rule shown. All are scored against R1/R2 with log score (bits), calibration, and the permutation gate.

Table 17: Quickstart task pack. Each row is a `task_type` with a fixed `answer_space` and a deterministic, pre-registered `resolution_rule` over later observable evidence. No subjective report resolves an outcome.

Task family	Answer space	Resolution rule (deterministic)
Recurring-commitment completion	{complete, defer, cancel, replace}	Calendar/task status or a pre-registered behavioural marker at the next scheduled occurrence fixes the label.
Meeting / event realization	{attended, no-show, rescheduled, cancelled}	Attendance signal (join event, location match, or logged presence) within the event window.
Response-latency bucket	{<1h, 1–24h, 1–3d, >3d, no-response}	Time from a flagged inbound obligation to the first outbound reply, bucketed; no reply before the horizon resolves to no-response.
Task continuation vs. switch	{continue, switch, pause}	The active task at T versus the active task after the next transition boundary; continuation iff the same task persists past the window.
Priority maintained vs. displaced	{maintained, displaced}	Whether the top-priority commitment at T still receives sustained allocation after the window, or is supplanted by a newly dominant one.
Engagement vs. avoidance of a defined obligation	{engaged, avoided}	Engaged iff a substantive action on the named obligation (reply sent, document edited, task advanced) occurs before the horizon; otherwise avoided.

Three product-facing instantiations

- **Note-taking / memory app (L0–L1).** Evidence = notes, tasks, calendar, and communication metadata already in the app. Natural tasks: recurring-commitment completion and response-latency bucket. Resolution is read from the app's own state at the horizon. No new sensing; the manifest declares tiers L0–L1 only.
- **Wearable audio recorder (L2).** Evidence = on-device transcripts and derived commitments/entities from ambient speech (Appendix M), not raw audio. Natural tasks: meeting/event realization and engagement vs. avoidance of a spoken obligation. Transcription stays local; only sealed predictions and resolved labels leave the device.
- **Multimodal glasses (L3–L4).** Evidence adds embodied context — objects, screens, documents, and locations in view. Natural tasks: task continuation vs. switch and priority maintained vs. displaced, where visual attention disambiguates transitions. Higher tiers carry higher sensitivity, so bystander redaction and on-device filtering (below) are mandatory, and

the evidence-tier ablation measures whether L3–L4 actually buys skill over the L2 audio-only condition.

Practical constraints (required for any real run)

- **Consent and eligibility.** Consenting adults only; per-instance `consent_status`/eligibility gating; ethics or IRB review before recruitment (Appendix E).
- **Bystander capture.** Visible recording notice where feasible; bystander redaction or exclusion; no capture in bathrooms, bedrooms, medical/legal settings, schools, children's spaces, or confidential meetings.
- **Missingness.** Missing evidence is a first-class state: instances without eligible evidence at T are reported, not silently dropped, and the R2 admission threshold (Appendix D) governs which strata are scored.
- **Device coverage.** Report coverage per target (fraction of the window with usable evidence); low-coverage targets are stratified separately, never pooled to inflate skill.
- **Local / federated.** Local-first storage; on-device filtering converts raw audio, video, and screen streams into task-relevant semantic events before persistence (Section 6.3).
- **No raw-data export.** Only sealed predictions, resolved labels, and aggregate metrics leave the client; no raw media or transcripts are exported by default.
- **Aggregate reporting.** Public output is the leaderboard row and aggregate diagnostics (Table 16); no per-target trajectories or raw inference objects are published (Section 13).

D Default R2 own-routine baseline specification

This appendix fixes a pre-registered default recipe (v1) for the R2 own-routine baseline. The recipe is provisional and protocol-configurable: the values below are defaults to be frozen before evaluation, not claims about any observed data. The intent is to make R2 a fair, routine-only competitor against which target-specific skill in bits can be measured without manufacturing lift against a weak baseline.

Admission. R2 is admitted for a given target and question type only when two conditions hold jointly: the target has enough prior resolved history (see minimum history below), and R2 beats R1 on a pre-seal validation slice (or a rolling historical criterion fixed before the seal). When R2 does not beat R1 on that slice, we report that R2 is not informative for that stratum and fall back to R1 as the reference; we do not manufacture target-specific lift by scoring against a deliberately weak own-routine baseline.

Minimum history. The default minimum is the stricter of 20 prior resolved opportunities or 14 days of eligible history per target and question type. Strata that do not meet this threshold are not admitted for R2 and are reported separately. The threshold is provisional and protocol-configurable.

Features (allowed). R2 may use only simple pre-prediction routine features: marginal historical frequency for the target $\times$ question type; recency-weighted frequency; persistence or previous target state; recent target-state transition counts; time-of-day; day-of-week; and known recurring calendar or deadline commitments that are available before prediction time.

Exclusions. R2 may not use any of the following: passive audio, video, or screen content; semantic text content beyond pre-specified routine labels; any future information relative to the evidence available up to time t; entrant model outputs; manually selected post-hoc features; or any representation learned from the scored test outcomes.

Fitting. R2 is organizer-fit and frozen before evaluation. Fitting is walk-forward only: no random cross-validation and no tuning against test outcomes. Parameters are estimated from history available before each prediction time and never revised using resolved test labels.

Smoothing. R2 applies a pre-registered smoothing scheme and a nonzero probability floor over the answer space A, so that every admitted answer receives positive probability and proper scores (log score in bits, Brier) remain finite. Exact smoothing constants and the floor are protocol parameters fixed before test results are seen.

Reporting. Skill in bits is reported against R1 and against R2 separately, never collapsed into a single number. Strata in which R2 fails admission are reported as such, with R1 as the reference. Incomparable answer spaces are not pooled without strata.

E Safe pilot configuration

The benchmark can be tested without assembling an uncontrolled lifelogging dataset. Table 18 lists a default minimal safe pilot configuration, and Figure 6 shows the federated, local-first architecture it assumes: raw capture stays on the participant device, and only sealed predictions, resolved labels, and aggregate metrics leave the client.

Table 18: Minimal safe pilot configuration. The benchmark can be tested without creating an uncontrolled lifelogging dataset. The following practical controls are the default for any pilot.

Default controls for a minimal safe pilot
• Consenting adults only.
• Local-first storage.
• No public release of raw video or audio.
• No raw third-party export by default.
• Only sealed predictions and aggregate metrics leave the device or client.
• Participant controls for review, pause, deletion, and opt-out.
• No capture in bathrooms, bedrooms, medical or legal settings, schools, children's spaces, or confidential meetings.
• Visible recording notice where feasible or required.
• Bystander handling as a hard constraint: on-device enrolled-speaker filtering with immediate discard of non-enrolled speech as the default L2 configuration; diarization uncertainty resolves to discard; egocentric video admitted only under venue-based capture rules [68,69,70].
• Short raw-media retention window unless explicitly consented otherwise.
• Ethics / IRB (or equivalent) review before recruitment.
• Special handling for third-party content, employee contexts, children, and sensitive data.

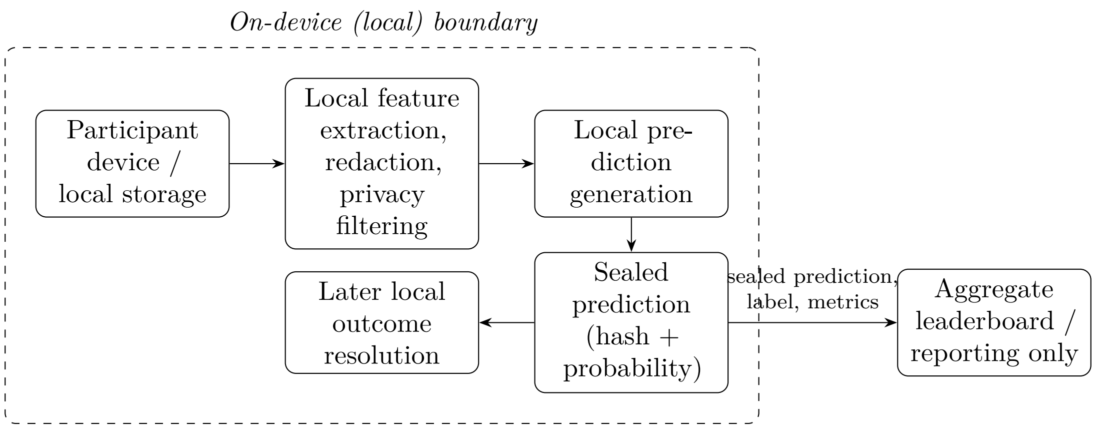

Figure 6: Safe federated pilot architecture. Raw audio, video, and screen data stay on the participant device: feature extraction, redaction, privacy filtering, prediction generation, sealing (hash and probability), and outcome resolution all run locally within the dashed on-device boundary. Only the sealed prediction (and, later, resolved labels and aggregate metrics) leaves the client; the off-device stage performs aggregate leaderboard and reporting computation only. No raw capture is transmitted.

F Limitations register

The full register of limitations summarized in Section 13, with mitigations where they exist. (L1) Latent targets are evaluated only through observable proxies. (L2) The federated, prospective design makes reproduction harder than a downloadable dataset and precludes third-party audit of sealing, labelling, and resolution, so results are holder-reproducible, not publicly auditable; we mitigate with a versioned protocol and shared harness. (L3) Single-instance and small-cohort evaluation limits external validity. (L4) Predictability is bounded [8] and heterogeneous, so reporting is per-instance. (L5) Capture can be reactive — being observed may alter the behaviour being predicted — so habituation windows and reactivity checks are required, and induced deviations must not be read as target-state transitions. (L6) Domain-generality is a property of the formulation, demonstrated in one domain; the cross-domain claim is specified, not established. (L7) The privacy and governance safeguards are design commitments, not implemented features: no pilot has run, no institutional review has been obtained, and the privacy-filtering layer is unbuilt. (L8) A calibrated behaviour predictor is dual-use, and the observe-not-intervene rule binds the benchmark, not downstream deployers. (L9) Longitudinal individual-level data carry selection and representation bias; a benchmark validated on a narrow cohort will not transfer, and per-instance reporting does not cure unrepresentative sampling. (L10) Ecological validity is limited: instrumented, consented capture may not reflect unobserved behaviour, and a habituation window does not guarantee it. (L11) Many target states admit annotation and label ambiguity — whether a commitment was 'abandoned' or 'deferred' can be genuinely contested — so resolution rules must be pre-registered and inter-rater reliability reported. (L12) Targets exhibit drift and non-stationarity: the dynamics a model fits can change, so R2 is re-fit walk-forward and skill decay is tracked rather than assumed stable. (L13) For some outcomes ground truth is intrinsically hard to establish — probabilistic, delayed, or only partially observed — which bounds achievable skill and is why scoring is proper rather than accuracy-based. (L14) Multimodal longitudinal capture is costly and burdensome to collect and store, constraining cohort size and biasing toward participants who can be instrumented. (L15) Results risk overfitting to a single person, household, or organization; the permutation gate guards against cross-target leakage but not against a model tuned to one idiosyncratic target, so external validity requires multiple independent targets. (L16) The benchmark measures prediction of behaviour, not understanding of intent: a high score is predictive alignment with observable future states, and any inference about goals, reasons, or inner life is explanatory audit (Section 6.2) subordinate to the sealed prediction, never established by the score.

G Participatory dynamics as auxiliary state variables

TargetSpace keeps epistemic scoring separate from target-state formation. Prediction accuracy is scored only against sealed external outcomes (Section 6). Separately, the harness can record observable feedback, attention, action, and constraint signals as *auxiliary* variables, letting researchers test whether adaptive feedback loops change the reachable target-state space and whether those changes improve predictions. Two processes must not be conflated: *epistemic resolution* is the observer's uncertainty falling as evidence accumulates — posterior concentration over the observer's maintained distribution; *participatory target-state formation* is a change in the target itself — through attention, action, feedback, and constraint — that alters which futures remain salient, reachable, reinforced, or stable. Posterior concentration in the observer is not target-state formation in the observed system. These proxies are a diagnostic layer only: the benchmark continues to score predictions solely against sealed outcomes, never against the auxiliary variables.

H Personal track: task specification and extended discussion

To show that TargetSpace instantiates runnable tasks and not only a framework, we specify one minimal personal-track task, the *Recurring-Commitment Completion Prediction.* It is synthetic and illustrative; no empirical results are reported.

- *Task.* Given timestamped evidence available up to a sealed time T, predict how a scheduled recurring commitment resolves by a later resolution time r.
- *Inputs (by tier).* Calendar and task history, communication metadata, and prior completion/defer/cancel history (L0); optional user-authored text traces (L1); optional passive-observation summaries (L4). See Section L.2.
- *Answer space.* $A = \{\text{complete, defer, cancel, replace}\}$.
- *Baselines.* R1 (population prior over the four outcomes) and R2 (the target's own recurring-commitment routine), both fit walk-forward on evidence before T.
- *Resolution.* A deterministic rule over later observable evidence — calendar status, attendance, task completion, or a pre-registered behavioural marker — fixes the outcome at r.
- *Scoring.* Log score (bits) and Brier against the sealed outcome; report skill over R1 and over R2 (Section 9).
- *Specificity.* A permutation test across matched targets should reduce or collapse the skill if a model's prediction was genuinely target-specific.

The task is small enough to run on current language models, agents, and retrieval or memory systems, and exercises the full stack: sealed prospective scoring, R1/R2 baselines, calibration, evidence tiers, and the permutation gate.

Horizon–abstraction hierarchy

The four task families are scored across the following horizon–abstraction hierarchy.

Table 19: A horizon–abstraction hierarchy for target-state prediction. Each level is scored independently; no single representation is assumed optimal across horizons, and resolution rules are deterministic or pre-registered. Uncertainty generally grows with horizon.

Level	**Horizon**	**Answer space / resolution rule**	**Baseline; metric**
Immediate action / next state	seconds–minutes	concrete next state; observed action	R2; log/Brier, calibration
Short-horizon task transition	minutes–hours	small discrete set; observed completion or switch	R2; Skill, top-k
Medium-horizon subgoal / commitment	hours–days	commitment set; follow-through within a window	R2; Skill, time-to-resolution
Longer-horizon priority / target-state shift	days–weeks	priority/regime set; sustained reallocation of attention/resources	R1/R2; transition-path likelihood
High-level regime / strategy transition	weeks+	coarse regime labels; observed regime change	R1; Skill, calibration

H.1 TargetSpace as a benchmark for human-centered world models

The world-model research direction defines intelligence operationally as the ability to predict the consequences of actions in an abstract representation space and to plan over those predictions [17,19]. TargetSpace applies that idea to human-assistive AI. The 'environment' is not a physical system but the latent state of a person: can a system infer, from passive observation up to a sealed time T and *before the user writes an explicit prompt*, the user's current target state, hidden constraints, and the likely next transition? The classical form is *state + action → next state*; the TargetSpace form is *passive longitudinal observation → belief state over latent targets* (Section 6.2) → *target-relevant future behaviour*, scored by calibrated skill over R1/R2 rather than by reward or reconstruction. Only the sealed prediction is scored — not planning, action, or the model itself; the target is abstract predictive state, not raw reconstruction, and the framing is architecture-neutral (Section L.1).

I Transition taxonomy and resolution rules

The full taxonomy of target-state transitions the personal track predicts (referenced from Sections 6.1 and 8.2) is given in Table 20. Each transition type carries a deterministic or pre-registered resolution rule grounded in later observable consequences, an ambiguity/evidence profile, and the baseline expected to be hardest to beat. The taxonomy is a provisional coding scheme, not an ontology: types are non-exclusive labels over episodes (a deadline-forced displacement is also constraint-forced), the scored object is always the answer space and its resolution rule, and type-assignment reliability is reported alongside outcome inter-rater reliability.

J Adopted components and extended related work

Table 21 itemizes the components TargetSpace adopts from prior work, their representative sources, and the use each is put to; the two anchors contributed here are the R2 own-routine baseline and the permutation specificity gate (Section 7).

Table 20: A taxonomy of target-state transitions the personal track predicts. Each has a deterministic or pre-registered resolution rule grounded in later observable consequences, not subjective report. "Likely baseline" is the baseline expected to be hardest to beat; "horizon" is the typical prediction range.

Transition type	**Operational description / resolution criterion**	**Ambiguity risk; evidence needed**	**Baseline; horizon**
Routine continuation	target persists; resolved by continued allocation/action matching the prior pattern	low; digital exhaust suffices	R2; short
Latent-but-active target	already pursued but unstated; resolved by later action or commitment	medium; behaviour + context	R2; short–medium
Emerging target	a new target becomes determinate; resolved by first commitment or resource expenditure after onset	high (onset timing); attention + context	R1/R2 weak; medium
Stabilization	a provisional target consolidates; resolved by repeated allocation across windows	medium; attention trajectory	attn-conditioned; medium
Competition	two targets contend; resolved by which receives sustained allocation	high; attention + outcomes	attn-conditioned; short–medium
Displacement / priority reversal	a new target supplants the prior one; resolved by a switch in sustained allocation	medium; attention + calendar/task	attn-conditioned; medium
Abandonment	a target is dropped; resolved by absence of follow-through past a window	medium; non-action over time	R2; medium
Resumption	a prior target returns; resolved by renewed allocation after a gap	high; longitudinal history	R2; medium–long
Constraint-forced transition	a constraint forces a change; resolved by the observed transition under the constraint	low–medium; constraint + outcome	R1; short
Opportunity-induced target	new evidence or opportunity creates a target; resolved by uptake after the encounter	high; evidence event + uptake	R1/R2 weak; short–medium

J.1 World models, predictive learning, and consequence prediction

A long lineage predicts the *latent* state rather than the surface: predictive coding [16,21]; world-model agents that plan over learned latent dynamics [15,20]; and joint-embedding self-supervised methods, notably LeCun's JEPA program [17–19], a *training framework* that predicts a target's *embedding* rather than reconstructing pixels or tokens. These reactive, predictive, and planning systems are candidate participants, not competitors; uniformly, though, they are compared by representation quality, reward, task success, or rollout quality — evaluations that do not themselves test calibration, skill over a target-specific routine, instance specificity, prospective sealing, evidence-value attribution, or multi-horizon target-state prediction. TargetSpace supplies that measurement layer and applies it to these classes on identical sealed prediction instances; what matters for scoring is whether the predicted transition resolves correctly against the external target, not how the prediction is represented internally.

J.2 Forecasting, person modeling, and goal recognition

Calibrated event forecasting is established — ForecastBench [13] — but scores external public events, not target-state transitions of a tracked instance, and uses no own-routine or permutation control. Person-modeling benchmarks are the nearest in framing but retrospective or uncalibrated: Park et al. [26] replicate individuals' survey answers (their own-consistency reference is the closest

Table 21: Prior components and TargetSpace-specific contributions. TargetSpace adopts prospective sealing, proper scoring, calibration, evidence ablation, and latent-predictive evaluation from prior work; its own contribution is target-specific prediction under strong controls — the R2 own-routine baseline and the permutation specificity gate — assembled around one evaluation question.

Prior component	**Representative sources**	**TargetSpace use**
Architecture-neutral world-model evaluation	AutumnBench / WorldTest [22]	Adopt; add calibration, R1/R2, permutation
Cross-architecture comparison	WorldPrediction [23]	Adopt; add proper scoring, sealing, specificity
Latent predictive learning	JEPA [17–19]	Evaluate as an architecture class
Proper scoring & calibration	Good; Brier; Gneiting & Raftery [10–12]; HELM [6]	Adopt directly
Sealed prospective prediction	ForecastBench [13]	Adopt
Evidence ablation	digital phenotyping / personal sensing [39,40]	Adopt; report as lift over R2
Goal recognition & resolution timing	goal-recognition design / WCD [14]	Adapt to sealed, tracked instances

cousin to R2, though used as a ceiling, not a baseline); KnowMe-Bench [32] formalizes person understanding as *retrospective* comprehension; EgoToM [33] infers a wearer's goals per clip by accuracy, without calibration or a persistent target; PersonaMem [28] tracks evolving preferences. Goal- and plan-recognition [14] infer latent goal posteriors under partial observability and define how early a target is identifiable, but the evaluation object differs: conventional goal recognition asks *which* goal, from a candidate set, best explains observed behaviour, whereas TargetSpace prospectively scores how a target's state, attention, and environment produce the *next* target-state transition (Table 20) under the R1/R2 and permutation controls. We adopt the resolution-timing idea for the metrics and re-cast it as a calibrated, prospective, instance-tracked measurement.

Table 22: The five-dimension conjunction across neighboring benchmarks (referenced from Section 12.2). Dimensions: 1: persistent individual / evolving entity; 2: longitudinal, continuously updated evidence; 3: strictly prospective / sealed prediction; 4: calibrated proper scoring with meaningful baselines; 5: target-state / meaningful-transition forecasting as the target. Labels are cautious judgments from each benchmark's design, not measured scores. Prior work contains important subsets; we are not aware of a benchmark combining all five in one prospective evaluation apparatus.

Benchmark / family	**1**	**2**	**3**	**4**	**5**
EgoToM [33]	absent	partial	partial	absent	arguable
ToMnet [27]	partial	partial	partial	absent	arguable
ForecastBench [13]	absent	absent	clear	clear	absent
PersonaMem [28]	clear	partial	absent	absent	arguable
Generative Agents [26]	clear	partial	absent	absent	absent
KnowMe-Bench [32]	clear	partial	absent	absent	absent
Goal / plan recognition [14]	partial	partial	arguable	absent	arguable
ARC-AGI [7]	absent	absent	clear	absent	absent
TargetSpace (this work)	**clear**	**clear**	**clear**	**clear**	**clear**

K Extended metrics and scoring details

This appendix carries the metric extensions and reporting details beyond the core scoring of Section 9.

Reporting and inference details

Predictions use a pre-registered nonzero probability floor; Skill is aggregated within homogeneous answer-space and horizon strata before pooling; Brier is a secondary robustness score, ordered answer spaces may add an ordered proper score, and no composite score is used. R1's exact fitting is pre-registered: population information only, leave-one-out with respect to the scored target, rolling, with smoothing for rare classes. Any recalibration uses a separate pre-seal split, is disclosed, and is never fitted on sealed outcomes. R2's exact feature set, recency kernel, minimum history, and admission rule are pre-registered; the default recipe (marginal frequencies, persistence, recent transitions, time-of-day and day-of-week effects, and recurring calendar/deadline commitments known before the prediction) is Appendix D, organizer-fit rather than entrant-fit, and never selected on test results. The allowed-feature boundary is operational, not rhetorical: the reference harness ships the frozen R2 implementation, and the admissible feature extractors are exactly those it enumerates — marginal and recency-weighted frequencies, persistence, time-of-day and day-of-week effects, and calendar-declared recurrences — so two compliant organizers produce the same baseline rather than materially different ones. Two admission caveats are pre-registered rather than discovered: stratum admission (R2 counted only where it beats R1) is a selection step, so admission is decided on a pre-seal slice with a multiplicity-aware rule and disclosed alongside results; and because rare transition strata give R2 little transition history to fit on, transition-stratum skill is reported only where R2 had pre-registered minimum support in that stratum, and is otherwise labelled unsupported rather than credited. The permutation gate swaps complete eligible histories rather than scrambling outcomes and is reported as an effect — true-instance Skill minus the mean permuted-instance Skill, as absolute and proportional skill loss with an interval and a pre-registered margin; its resolving power grows with the number of instances permuted, so it is directional and operational in a five-person pilot and a powered test only above a stated minimum cohort size. Uncertainty respects the repeated, serially dependent structure of predictions within a target: within-person differences use day-blocked analysis and between-person summaries use person-clustered (cluster-bootstrap) intervals, since the person — not the individual prediction or day — is the independent unit for between-person claims; the current reference implementation reports point estimates. Calibration is assessed primarily through the calibration intercept and slope (logistic recalibration parameters) and a reliability decomposition; top-label ECE is reported only as a secondary diagnostic under a fixed pre-registered binning scheme with a bootstrap interval, since ECE is positively biased and sensitive to bin count and placement at the per-stratum sample sizes of a five-person pilot, and its pass/warn/fail bands are therefore interpreted as diagnostic bounds rather than a hard gate below a pre-registered minimum stratum size.

Multiscale consistency (proposed components)

Because levels are scored separately, we also specify cross-level checks as future evaluation components, not results: *cross-horizon consistency* (a detailed prediction should not contradict the abstract target state without explicit uncertainty or branching); *refinement* (as the horizon shortens and evidence arrives, abstract predictions should sharpen into concrete ones); *subgoal coherence* (predicted intermediate states should form a plausible path to the higher-level target); *horizon-conditioned calibration* (confidence is evaluated separately at each horizon); *hierarchical abstention* (a system may abstain at the detailed level while keeping a calibrated abstract prediction); and *rollout drift* (repeated one-step predictions may diverge from external reality even when each step is locally accurate).

K.1 Metrics for possibility-space resolution

Predictive lift by evidence tier is Skill computed as a function of the evidence rung — the evidence-ablation read-out; no new mathematics. **Top-*k* target-state recall** complements distribution-level Skill for large answer spaces. **Transition-path likelihood** is the length-normalized likelihood of the realized path of target-state transitions, in bits vs R1/R2 — the natural metric for the transition track. **Time-to-correct-resolution** and **false-resolution rate** measure *when* a model resolves the possibility space: respectively, how early its distribution concentrates correctly on the realized target, and how often it concentrates confidently and early on the *wrong* one. We are explicit that 'how early can a target be identified' is **not** a new idea: it descends directly from goal-recognition design and worst-case distinctiveness (WCD) [14] and from online goal recognition (recognition time, false-positive rate). Our contribution is to operationalize it as a *proper-scored, calibrated, prospective* quantity on tracked instances, paired so that a model must achieve early correct resolution *and* low false resolution together. We freeze and unit-test these definitions before claiming them and report none as results pre-pilot.

Table 23: Metric provenance. The honesty rule: claim the conjunction and the R2 + permutation anchors; concede the rest as adopted, adapted, or (for resolution-timing) a new operationalization of an existing idea.

Metric	**Status**	**Nearest prior art (cite, do not claim)**
Skill / lift / entropy reduction	adopted (information gain)	skill score; Gneiting & Raftery [11]
Log + Brier, ECE gate, sealing	adopted	[10,12,13]; calibration-as-axis [6]
R2 own-routine + permutation gate	**TargetSpace anchor**	own-history baselines & self-consistency as cousins
Top-k recall; transition-path likelihood	adapted	retrieval; cell-fate / trajectory likelihoods
Time-to-correct-resolution	operationalization, not new idea	goal-recognition design / WCD [14]; online goal recognition
False-resolution rate	new metric (concept has precedent)	premature-commitment / false-stop / FPR in goal recognition

L Evaluation grid, evidence ladder, and tracks

This appendix specifies the evaluation grid summarized in Section 7: the architecture classes and the person-domain evidence ladder with its privacy-risk taxonomy.

L.1 Architecture classes

The grid does not rank architecture classes in the abstract. Systems become comparable only because each one emits, or is adapted to emit, a calibrated probability distribution over the *same* answer space A on the *same* sealed instance, scored by the same proper rule. A reactive policy, an LLM, a VLM, a JEPA-style latent-predictive model, a symbolic or probabilistic system, a hybrid, and a human self-report are all evaluated at the level of the prediction output, not by inspecting internal representations. Because the output adapter — which maps a system's native output (action, token sequence, latent prediction, query result) to a distribution over A — can change measured performance, it is disclosed and reported as part of the system under test: it must be specified before the prediction is sealed, so reported skill reflects the system as actually run rather than a post hoc reinterpretation.

Table 24: Architecture classes. LLMs and JEPA-style models are complementary components, not rivals: TargetSpace asks whether *any* of them produces calibrated, target-specific predictions. Human self-report and the oracle anchor the achievable range and are not ranked. The third column indicates what each class supports natively versus through a disclosed output adapter; labels are cautious, classes are not ranked, and real systems are often hybrid.

Architecture class	**What it is**	**Prediction interface (native / via adapter)**
LLM-only	text-in, distribution-out language model [5,6]	distribution native; explicit state model via adapter
VLM	vision-language model over image / video	distribution via adapter; no native state model
Multimodal agent	tool-using agent with retrieval and memory [47]	distribution via adapter; target-specific memory possible
JEPA-style / latent predictive (incl. world models)	learns latent dynamics, predicts in representation space [17–19]	state and action-conditioned prediction native; calibrated distribution via adapter; planning by search
Symbolic / probabilistic	explicit structured or Bayesian model [46]	calibrated distribution and state prediction native; planning architecture-dependent
Reactive / imitation policy (incl. VLA)	maps observation + instruction to action [24]	action native; consequence distribution via adapter; not intrinsic
Hybrid	any combination, e.g. policy + learned subgoal or world model [20,24]	architecture-dependent; reactive and predictive components may coexist
Human self-report *(baseline)*	the target, or an expert, predicts itself [13,37]	distribution via elicitation; not ranked
Oracle upper bound *(baseline)*	a privileged predictor; approximate ceiling (benchmark construct, no external referent)	not ranked

L.2 The evidence ladder

The evidence ladder is the benchmark's measurement instrument. Tiers are cumulative and ordered by increasing capture intrusiveness (and, roughly, privacy sensitivity), *not* by assumed predictive power. By reporting target-specific skill (over R2) as a function of tier, the ablation turns "does richer observation add target-specific information?" into a measured quantity. We give the canonical *person-domain* ladder; the apparatus is domain-general wherever a strong own-routine R2 baseline and deterministic resolution rules exist, though only the synthetic personal track is instantiated.

The ladder also encodes a distinction richness alone does not (developed in Section 5): lower tiers (L0–L1) are records the target already produced, beginning *after* relevance was assigned and largely encoding routine, whereas higher passive tiers are **pre-interpretive** — captured before the

Table 25: The person-domain evidence ladder. Tiers are ordered by increasing capture intrusiveness (and, roughly, privacy sensitivity) — from already-externalized digital residue toward pre-interpretive observation; they are *not* ordered by assumed predictive power, which is exactly what the ablation measures.

Tier	Evidence (person domain; operational examples)	Primary sensitivity risks
L0	low-content behavioural metadata: calendar timestamps, communication metadata, app/event logs, coarse routine traces	social-graph and rhythm inference
L1	user-authored text: notes, chats, documents, search/query text, task lists	third-party content; workplace/confidential exposure
L2	speech and transcript layer: audio recordings, ASR transcripts, speaker turns, optional prosody	bystander capture
L3	screen and interaction context: screen recordings, UI activity, browser/app context, document-interaction traces	workplace/confidential exposure; intimate-behaviour inference
L4	egocentric / passive multimodal observation: wearable or scene video, ambient audio, images, environmental context	bystander capture; re-identification
L5	location, mobility, and physical-world traces: GPS, Wi-Fi/Bluetooth proximity, visits, commute patterns	location-trace risk; re-identification
L6	physiological / biometric / specialized sensors: heart rate, sleep, gaze, motion, wearable health signals	health and biometric inference

target fixed what would matter, they carry **retrospective option value** (the same recording can be re-scored under questions chosen later) and are, more precisely, *independent from retrospective human interpretation* rather than "objective". The grid is the cross-product: each leaderboard cell names an (evidence tier, architecture class) pair, with R1, R2, human self-report, and the oracle spanning all tiers. Because the relevance of an observation may only become clear retrospectively, TargetSpace treats raw-evidence preservation, provenance, and ablation as first-class concerns: it can ask not only whether a system used evidence but *which* evidence was necessary to improve a target-specific prediction. Preservation is always bounded by the consent, federation, aggregate-only, and retention limits of Section 13; the benchmark scores evidence value, it does not license indiscriminate retention. Retrospective option value is bounded by purpose limitation: re-scoring happens only under a registered question set and within fixed retention deadlines, so preservation never shades into indefinite retention [60].

M Industry case cluster: from AI recorders to first-person memory systems

This appendix gives the analytical survey behind the motivation of Section 1. It is evidence that the capability class is arriving, not empirical validation of TargetSpace, and not a claim that any vendor misuses data; each such system also has clearly beneficial uses (accessibility, memory support, productivity, care coordination, safety).

Three product groups are shipping the capability class: always-on ambient audio recorders, wearable "memory" devices that persist and summarize conversation, and first-person audiovisual glasses that add embodied visual context [49–55]. Across all of them the trajectory is consistent, and the TargetSpace-relevant observation is common: product value is increasingly located in the

Table 26: Evidence substrates for the personal track and their role in TargetSpace (referenced from Section 5). Self-report is retained as an auxiliary channel and a baseline, not discarded; passive capture is not claimed unbiased, only to carry externalized, measurable, partially normalizable bias. Tiers refer to the ladder above; “standardizability” is how readily the bias can be characterized and normalized across people and time.

Substrate	**What it supplies / characteristic bias**	**Standardizability of the bias**	**Role in TargetSpace**
Self-report / stated goals	declared beliefs, goals, feelings; bias endogenous — selective, post-hoc, socially filtered, introspection-limited	low; varies across people and over time	auxiliary channel + human-self-report baseline
Digital exhaust (L0–L1)	calendars, messages, clicks, notes; already-externalized, routine-heavy residue	moderate; well-defined logs	primary R2 signal; ablation floor
Passive audio (L2)	speech, prosody, turn-taking, timing; ASR error, mic placement/range, bystander capture	device-characterizable (WER, coverage)	candidate lift over exhaust/self-report
Passive audiovisual / egocentric (L3–L4)	embodied context: what is seen, where, which objects/people; FoV, occlusion, framing	device-characterizable; higher sensitivity	candidate lift; tested by ablation
Sensors & future internal-state proxies (L5–L6)	location, mobility, physiology; prospectively affective/internal-state proxies; sensor-specific error	sensor-calibratable	candidate lift; still scored on observable outcomes

derived, longitudinal layer — summaries, entities, commitments, routines, persistent memory — not in the recording itself. That is exactly the layering of Table 27, and the same ladder is climbed with different signals by assistants and agents over email, calendar, and documents, by enterprise copilots, tutors, care systems, and recommenders — any system that converts accumulated longitudinal traces into a persistent, future-relevant model of a particular target (Section 6).

Table 27: From capture to target-state inference (referenced from Sections 6 and 13). Each layer is more compressed, searchable, and actionable than the one below; TargetSpace evaluates, and the governance discussion concerns, the top layers, not only the raw signal.

Layer	**Representation**	**Example capability**	**Primary risk**
Signal	raw audio, video, screen, location, sensor traces	record or observe events	unauthorized or unnoticed capture
Text / perception	transcripts, OCR, visual labels, logs	make events searchable	searchable bystander speech and activity
Structure	speakers, entities, topics, tasks, commitments	organize experience into relations and obligations	relationship and obligation mapping
Memory	persistent longitudinal profile, embeddings, summaries, routines	retrieve and personalize over time	pattern-of-life inference
Target state	predicted goals, needs, actions, future commitments	anticipate (or steer) future behaviour	manipulation, surveillance, unwanted steering