TargetSpace: Benchmarking Personal Intelligence by Target-Specific Forecasting Under Partial Observation

A prospective framework with own-routine baselines, permutation specificity, calibration, and evidence ablation

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

Abstract

AI systems increasingly claim to *know*, *remember*, *personalize to*, and *act on behalf of* particular users — assistants, agents, companions, long-term memory systems, coaches, and adaptive copilots. Yet current evaluations cannot establish whether such a system has formed a durable, calibrated, prospectively useful model of the individual, as opposed to retrieving stored facts, replaying remembered preferences, imitating style, or exploiting short-horizon behavioural correlation. We call the missing capability **personal intelligence**: the capacity to form, update, and prospectively test a calibrated model of a specific person over time — their changing priorities, latent goals, constraints, likely transitions, and uncertainty. The premise is that a person is not a profile to retrieve but a **lived trajectory** to forecast — an unfolding, situated stream of attention, context, constraints, and emerging goals from which the next action follows — inferred only from observable longitudinal traces, never from stored facts or stated preferences alone. We introduce a prospective benchmark for it. The benchmark scores sealed, proper-scored forecasts of a person's future **target-state transitions** under **partial observation**, certified by a stack of controls: a population-prior baseline (**R1**) the model must beat to exceed base rates; a strong, person-specific **own-routine baseline (R2)** it must beat to exceed routine and memory replay; a **calibration** gate; and a **permutation specificity test** (a model's forecasts for one person, scored against another person's outcomes, must lose their skill). An **evidence-tier ablation** measures whether passive observation adds information beyond digital exhaust, and an **architecture-class** grid compares language, vision-language, latent-predictive, symbolic, memory, and human-self-report systems on identical sealed instances. Crucially, some target states are not fixed labels but form through attention and action; we therefore treat attention as both evidence and a candidate process in target-state formation, and test whether attention trajectories add forecasting skill beyond routine and stated goals. We do not claim forecasting exhausts understanding; self-report and observation are complementary channels whose relative value is empirical (**asymmetric observability**). Personal intelligence is the flagship track of **TargetSpace**, a framework that also extends, by stated criteria, to other tracked-target domains. Proper scoring, calibration, sealing, evidence ablation, and architecture-neutral evaluation are adopted from prior work and cited; the contribution is their **conjunction**, anchored on the own-routine baseline and the permutation specificity gate. This is a pre-pilot proposal; no empirical results are reported, only the personal track is implemented (synthetic), raw data is never public, and the benchmark measures predictive skill, not licence to act on it.

1 Introduction

Modern AI products increasingly claim to understand a *specific person*: assistants that remember you, agents that act on your behalf, companions, long-term memory layers, coaches, and adaptive copilots all rest, implicitly or explicitly, on the promise that the system has learned something durable about *this* individual and can anticipate what they will do or need. The promise is valuable, commercially central, and almost entirely unaudited. The same structure recurs wherever a system claims to model a particular evolving entity — an organization, a software agent, a market, a patient, a biological specimen — but the personal case is the most legible and the most consequential, and we take it as the flagship.

Underlying the personal case is a stronger ambition than personalization usually names. We are not trying to model what a person *knows*. In a precise and deliberately bounded sense, we are trying to model what it is like to be that person[1] — not their inner experience, to which a system has no access and about which we make no claim, but the **lived trajectory** from which their behaviour emerges: the unfolding, situated stream of attention, context, pressure, memory, constraint, and possibility that shapes what becomes salient and what they do next. A person is not a profile to retrieve, a preference list to replay, or a demographic vector to look up; a person is a trajectory in motion. TargetSpace asks whether a model can infer that trajectory from observable longitudinal traces well enough to forecast its next turn — a question the rest of this paper makes entirely operational.

The problem is that **many forecasts that look like understanding are not target-specific at all.** A model can appear skilled by predicting what *usually* happens: most planned meetings occur, most emails are eventually answered, most projects slip a little, most regenerating tissues grow back. It can appear personalized by replaying a target's *routine*: this person checks email at 9am, this project ships on Fridays, this robot usually completes the grasp. Both are useful, and both can produce a confident, plausible forecast — while revealing nothing about whether the model has captured anything specific to *this* target beyond population base rates and obvious routine. Retrospective explanation does not settle the question either: a model can rationalize what already happened without ever being held to what happens next.

Personal intelligence versus its cheaper substitutes. The personal claim has especially many ways to look satisfied without being: storing personal facts, retrieving over conversation history, replaying remembered preferences, optimizing recommendations, imitating a person's style, or fitting short-horizon behavioural correlations. Each is useful and none is, by itself, understanding the person. We reserve **personal intelligence** for the stronger capability: the capacity to form, update, and prospectively test a calibrated model of a specific individual over time — their changing priorities, latent goals, constraints, likely transitions, and uncertainty. The motivating gap is concrete: *a substantial class of AI products claims to remember, personalize to, adapt to, or act on behalf of users, yet the field lacks a rigorous prospective benchmark for distinguishing durable person understanding from retrieval, preference storage, stylistic imitation, and short-horizon behavioural correlation.* The benchmark is designed to give builders a credible answer to one question — *does the system actually improve its calibrated model of a particular person over time?* — without equating commercial activity with scientific importance, or the number of personalization products with a solved problem.

There is a second way a forecast can look like understanding without being it. Modern systems

[1]Here we borrow the phrase "what it is like" only as a motivating contrast between third-person description and situated experience, following Nagel's classic formulation [92]. TargetSpace does not claim access to consciousness, qualia, or private subjective states; it evaluates only whether models can infer future target-state transitions from observable longitudinal traces.

increasingly impress because they *generate* plausibly — fluent explanations, convincing simulated behaviour, smooth robot motions. But generation is not agency. Reliable long-horizon agency is generally strengthened by anticipating the *consequences* of actions and events — how a specific target's state will transition — under partial observation. A system that produces a plausible next action without predicting the state it leads to cannot be held to an outcome, and action success alone does not establish that it can explicitly forecast or compare future consequences. TargetSpace evaluates this missing layer: not whether a system can *say* or *do* something plausible, but whether it can forecast what a specific target will *do or become* next.

TargetSpace asks not whether a model can say what usually happens, but whether it can forecast what this target will do or become next, under specified evidence constraints.

The right question, then, is comparative and prospective: can the model forecast this target system better than (i) **the average system** — the population prior — and (ii) **the target's own routine**? Only a model that beats both, stays calibrated, and loses its skill when its forecasts are matched to the *wrong* target has demonstrated something specific to this system rather than to its crowd or its habit. **TargetSpace** operationalizes exactly this. It is a benchmark framework for prospective, sealed, proper-scored forecasting of a target system's future **target-state transitions** under **partial observation**, with the controls that make target-specificity measurable: a population-prior baseline (**R1**), a strong instance-specific own-routine baseline (**R2**), a calibration gate, a permutation specificity gate, and an evidence-tier ablation. We name these the *target-specificity stack* (Section 3).

TargetSpace is not a benchmark framework for whether a model can predict what usually happens. It is one for whether a model can predict what this target will do next. Personal world modeling is the motivating first application, because understanding individual people is valuable and is currently underserved: today's personal AI is trained largely on *digital exhaust* — calendars, messages, clicks — which often overrepresents already-externalized decisions and recurrent behaviour. But the formulation applies to any target system for which repeated observations and resolved outcomes make an own-routine baseline and a permutation test well-defined. We therefore describe TargetSpace as **domain-general in formulation** and **instantiated first in personal world modeling**, and we make no claim of cross-domain empirical validation in v1.

The closest empirical neighbours to this proposal are the experience-sampling and ecological-momentary-assessment tradition [61], personal-sensing work [62], and digital phenotyping [64], which likewise gather longitudinal, individual-level signals from everyday life. TargetSpace differs in what it places under evaluation: rather than predicting outcomes per se, it makes target-specificity itself the measured quantity, through the R1 population-prior and R2 own-routine baselines, proper scoring, calibration, sealed forecasts, and the permutation specificity test. The aim is complementary to these lines of work — we adopt their concern with situated, person-level data while shifting the object of measurement from prediction accuracy to whether a forecast is specific to a particular target rather than to a generic prior.

1.1 Understanding is the capability; forecasting is the measurement

Understanding is a capability. Forecasting is a measurement.
The target is not a profile. The target is a lived trajectory.

The capability we care about is understanding a specific system: an accurate, continuously updated grasp of how its situation and dispositions evolve. It is latent and impossible to score directly. Forecasting is its measurement — the observable quantity a system possessing the capability should

produce — the role classification played for ImageNet [1] and standardized tasks have played for language and code [20,21]. Across the sciences, predictive success is how a model is judged when the thing modeled cannot be inspected: weather theories by proper-scored forecasts [18,22], market and behavioral models by out-of-sample prediction [14,15]. We do not claim prediction *is* understanding; we claim it is its strongest available operational signal — and, crucially, only when the prediction is shown to be about the *target* and not the crowd. TargetSpace is therefore **forecast-first, not forecast-only**: it establishes whether a system has target-specific predictive skill before asking what structure, if any, explains it, and a good score is a precondition for an explanatory claim, never proof of one. We separate four things a model might do — *retrospectively describe* a target, *infer its current state*, *predict its next action*, and *forecast its next target-state transition* — and score only the last. A model can describe a person accurately yet fail to forecast the next transition; TargetSpace treats that failure as evidence that the model has not captured the relevant evolving dynamics, not merely that it lacks facts.

1.2 The observation bottleneck, and possibility-space resolution

Why is this hard, and why now? The binding constraint is an **observation bottleneck**: we rarely see a system's latent state, only a thin, biased residue of it. For a person that residue is digital exhaust, produced *after* the person already decided what was worth recording; for a cell population it is a handful of assays; for an organization, filings and releases. A model that learns only the surface regularities of this residue learns a routine, not a generator — which is exactly the failure target-specificity is meant to expose. What should a forecast be *about*? A partially observed system at any moment occupies a **possibility space**, a distribution over the states it might next move into, and an adaptive system does not wander it at random: it acts to reach certain **target states**. Understanding the system is operationalized as forecasting how its possibility space **resolves** toward those target states, and especially the **transitions** between them — the moments when *what the system is acting to reach* changes, where routine breaks and target-specific skill is decisive. Two processes must not be conflated here. **Epistemic resolution** is the *observer's* uncertainty falling as evidence accumulates — posterior concentration over a distribution the observer maintains. **Participatory target-state formation** is a change in the *target itself*: through attention, action, feedback, and constraint, an adaptive system alters which futures stay salient, reachable, reinforced, or stable. Posterior concentration in the observer is not target-state formation in the observed system, and the benchmark keeps them apart by scoring forecasts against sealed external outcomes (Section 5.2) — a model earns credit for anticipating how the target's own process unfolds, not for growing more confident. We use *resolve* only as shorthand for a distribution concentrating, and make no physical claim.

Compressed artifacts versus the process behind them. The residue is not only thin; it is the wrong *kind* of trace. Current systems are trained and evaluated primarily on artifacts people intentionally produce — text, code, clicks, documents, messages, task logs. These are *compressed* traces of cognition and action: they record the outputs a person chose to externalize, not the continuous process that produced them, and they appear only *after* intent has already been summarized or formalized. They therefore omit the observational stream from which goals, constraints, interruptions, attention, and state transitions could be inferred as they happen. We argue that the hard missing data layer for a human-centered world model is therefore not merely a better model architecture but the absence of a **longitudinal, passive-observation corpus of real activity over time, aligned to goals, decisions, task transitions, and outcomes**. TargetSpace targets this missing layer by specifying an evaluable prediction target over such observation; it does not claim that such a corpus exists, and Section 8 is explicit about what would

have to be built and why raw capture alone is insufficient.

Memory as a forward-looking belief, not an archive. Personal AI inherits a tempting but mistaken default: treat memory as storage and success as faithful recall. Human memory is not built that way. Autobiographical memory is reconstructed in the service of the rememberer's current goals rather than replayed verbatim [81], and the constructive system that reassembles the past is the same one that simulates possible futures — memory is selective and prospective by design [82]. Its value is the preservation of *behaviourally relevant* structure — what mattered, what changed, what remains unresolved, and what the person is now oriented toward — not fidelity to the record. TargetSpace adopts this functional stance: it does not reward a system for reproducing a person's logged history (the recall a retrieval system already supplies), but for converting passive longitudinal observation into a compact, continually updated estimate of the person's latent targets that improves prediction of their future target-state transitions. A benchmark scored on recall measures the archive; TargetSpace is scored on the forecast.

1.3 Attention and participatory target-state formation

Attention is limited and selective: it governs which information is sampled, represented, rehearsed, learned, and turned into action [71,72]. This has a consequence the benchmark must take seriously. In adaptive systems, target states are not always fixed hidden labels awaiting decoding; some *form* over time. Repeated allocation can reinforce and stabilize a provisional objective; withdrawal can weaken, delay, or displace one; and newly encountered evidence or opportunity can make determinate a target that previously was not [72,73]. Goals and attention are reciprocal — goals guide attention, and attention in turn helps maintain, revise, and sometimes constitute practical objectives [73] — a dynamical loop rather than a one-way law:

$$z_t \;\rightarrow\; a_t \;\rightarrow\; (e_t,\, u_t) \;\rightarrow\; z_{t+1},$$

where z_t is the latent target state, a_t attention allocation (or an observable proxy), e_t newly encountered evidence, and u_t action. Attention is therefore both *evidence* of the current target state and one *process* through which the next is formed. We claim no more than this cognitive-dynamical reading: it is not a physical, quantum, or consciousness claim; attention does not create targets from nothing; and prior dispositions, habit, environment, and constraint remain primary. The loop is, however, measurable prospectively — which is what the benchmark exploits, asking whether attention trajectories add forecasting skill beyond routine and stated goals (Sections 7.2, 8.1). The benchmark does not require the strong claim that attention causes target formation; it requires only the testable claim that attention trajectories may improve prospective prediction of target-state transitions. Attention may instead proxy unobserved causes, so establishing a causal role would require intervention, natural experiments, or explicit structural assumptions; the design treats attention as a candidate predictor evaluated by proper scoring, not as a mechanism.

1.4 Contributions

(1) Target-specific forecasting as a well-posed evaluation object. We distinguish three things often conflated — *generation* (predicting surface outputs), *imitation* (predicting plausible actions from demonstrations), and *target-state forecasting* (predicting the consequence, i.e. the latent target-state transition, of actions and events) — and define and score the third, under partial observation, distinguished from predicting population base rates, a target's routine, next-action, or external events (Sections 2, 4–5). **(2) The target-specificity stack.** A stack of controls centered on a strong instance-specific **own-routine baseline (R2)** and a **permutation specificity gate**,

while adopting proper scoring, calibration, sealing, and evidence ablation from prior work (Section 3). **(3) An evidence-tier × architecture-class grid** for comparing LLMs, VLMs, JEPA-style models, multimodal agents, symbolic/probabilistic systems, hybrids, human self-report, and oracle bounds on identical sealed instances (Section 6). **(4) A first instantiation in personal world modeling**, where passive first-person evidence is tested against digital exhaust for target-specific predictive lift (Section 8). **(5) Cross-domain extension criteria** for other target systems, stated as formulation rather than as validated empirical breadth (Section 9). Throughout, the grid scores *consequence prediction* and target-state forecasting independently of whether a system is a reactive policy, a predictive world model, an explicit planner, or a hybrid (Sections 5.4, 6.1). The audited novelty is the *conjunction* assembled around one question, not any individual ingredient (Sections 3, 11).

TargetSpace's distinctive contribution is not a new scoring rule, model architecture, or dataset; it is a prospective target-specificity test, in which each sealed forecast over an answer space A at evidence time t is evaluated for skill in bits over a population prior (R1), skill over the same agent's own-routine baseline (R2), calibration, and collapse under target permutation, all reported by evidence tier. The contribution is the conjunction of these controls applied together to a single prospective forecast, not any one of them in isolation. Of the four, two are not already standard practice and anchor the design: the R2 own-routine baseline, which measures whether an agent improves on its own habitual behaviour rather than only on a generic prior, and the permutation specificity gate, which tests that a reported skill score collapses when forecasts are matched to the wrong targets. The remaining controls — proper scoring against R1 and calibration — make the test interpretable, but it is the combination, with these two anchors, that distinguishes the proposal.

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)
Personal world modeling / personal intelligence	Specified as flagship track
Synthetic demonstration harness	Implemented; sanity check only
Human pilot results	Not claimed (no pilot run)
Cross-domain empirical validation	Not claimed
Passive observation improves forecasting	Hypothesis; to be tested by evidence ablation
Attention causes target formation	Not claimed; tested for incremental predictive value
Public raw dataset	Not provided; not planned
Deployment legitimacy / permission to act on forecasts	Not claimed

2 Why Target-Specificity Matters

This section makes the central problem concrete, because it is what TargetSpace exists to measure and what most current evaluations cannot distinguish. A base rate is the average trajectory; a routine is a profile of the *typical* trajectory; target-specificity is the demand that a model track *this* target's trajectory precisely where it departs from both. **Generic prediction can masquerade as understanding.** Two cheap strategies produce convincing forecasts without any target-specific knowledge. The first is learning **base rates**: a model that knows the population statistics — how often meetings occur, emails get answered, projects slip, grasps succeed — will look calibrated and often correct, while knowing nothing about any individual target. The second is learning a **routine**: a model that replays a particular target's recent regularities will look *personalized* while having

captured only habit. A benchmark that scores raw accuracy or agreement cannot tell either strategy apart from genuine understanding.

Target-specific understanding is what remains after both are subtracted. Operationally it means three things together. First, the model must **beat a strong own-routine baseline** (R2): predicting a target's habit is exactly what R2 already does, so skill over R2 is skill the routine does not contain. Second, the skill must be **calibrated**: a confident lucky streak is not understanding. Third, and most distinctively, the skill must **depend on the correct target-instance pairing**: if we permute outcomes across targets — scoring the model's forecasts for target i against the realized outcomes of a different target j — genuine target-specific skill should *collapse*. If it survives permutation, the model was modeling task or population structure, not the target.

Four examples make the distinction tangible. **Person.** Predicting that someone will eventually answer an email is base-rate prediction; the target-specific question is whether the model knew *this* person would ignore *this* message because their attention had shifted to another project this week. **Organization.** Predicting that a project will slip is routine; the target-specific question is whether the model knew *this* project would slip *beyond this organization's normal delay pattern*. **Robot.** Predicting task success from population success rates is generic; the target-specific question is whether the model captured *this* environment's or instance's particular constraints. **Biological tissue.** Predicting species-level regeneration is a base rate; the target-specific question is whether the model inferred *this* specimen's latent target-state response to a specific perturbation. In every case the gap between the generic and the target-specific answer is exactly what R2 and the permutation gate are built to expose.

3 The Target-Specificity Stack

We present the controls not as a loose conjunction of desirable properties but as a **stack**: each layer removes a way a forecast can look like understanding without being target-specific. TargetSpace adopts most of these tools from prior work (Section 11); its contribution is assembling them around a single question — *is the forecast genuinely about the target system?* — and adding the two layers that directly test it.

The stack composes into a single decision rule (Figure 4): a forecast earns *target-specific* credit only if it beats R1, beats a strong R2, passes calibration, and *fails* under permutation. Removing any one layer reopens a way to score well without target-specific understanding, which is why we present them together rather than as separable contributions.

4 From Goals to Target States

The object the stack scores is a target system's **target-state transitions**. The intuitive version — forecast a system's *goals* — smuggles in conscious intent, which excludes most adaptive systems. 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 — the minimal generalization that lets one apparatus span persons and non-persons.

Target states need not presuppose conscious, verbally reported goals. In the personal track the claim is operational: a target state is a *behaviourally resolved* configuration the person tends to reach, maintain, resume, or abandon — a commitment, priority, or task regime. Cross-scale biological examples — e.g. the target morphology a regenerating organism rebuilds toward [25] — motivate this broader, mind-agnostic vocabulary but are not required: the personal benchmark stands on behavioural forecasting alone, and nothing here is a metaphysical claim about tissues

Table 2: 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.

Layer	What it removes / certifies	Status
1. Prospective sealing	forecasts sealed and timestamped before outcomes exist — removes hindsight rationalization and contamination	adopted [2,28]
2. Proper scoring	log / Brier scoring of the whole distribution — rewards calibrated probabilities, not cherry-picked accuracy	adopted [14,15,18]
3. R1 population-prior baseline	skill must exceed population base rates — filters out generic base-rate prediction	adopted
4. R2 own-routine baseline	skill must exceed the target's *own* strong routine — filters out routine mimicry	**anchor (this work)**
5. Calibration gate	expected-calibration-error (ECE) threshold — filters out overconfident lucky prediction	adopted [43]
6. Permutation specificity gate	skill must collapse when forecasts are matched to the wrong target — tests dependence on the correct instance pairing	**anchor (this work)**
7. Evidence ablation	skill as a function of evidence tier — measures which streams add target-specific information	adopted (e.g. active/passive sensing)

having goals. (We keep two researchers distinct: *Michael* Levin, developmental biology, motivates target states beyond conscious intent; *Sergey* Levine, reinforcement learning, is a separate line of work.) *Transitions* matter because the moments when what a system is acting to reach changes are where non-routine behavior originates and where a routine baseline fails — the regime in which target-specific skill is decisive. A transition is not a single event type: a target may newly emerge, stabilize, persist, compete with another, be displaced, abandoned, or resumed, or be forced by constraint or created by opportunity. TargetSpace treats this typology as a forecasting target with deterministic resolution rules (Table 6), which is what separates it from recognizing *which* fixed goal is currently active (Section 11.3). The path a target traces through these transitions is what we call its **lived trajectory**: not a fixed profile of what it is, but the unfolding sequence of what it is oriented toward. Forecasting the target means anticipating that sequence — which state becomes salient next — from observable evidence; it is this, and no claim about inner experience, that the benchmark scores.

5 What TargetSpace Is, and Is Not

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 9). 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 from any agent that solves its issues.

5.1 Definitions and the forecast unit

A **target-system instance** is a specific partially observed adaptive system tracked over time (a person, animal, robot, tissue, organization, project, software agent, market). Its **current abstract state** is the latent configuration generating its behaviour at time t; a **latent target state** is a

configuration it acts to reach or maintain; a **transition path** is the sequence of target states it moves through — the formal counterpart of the *lived trajectory* of Section 4, and the object a model is asked to forecast. A benchmark instance is the tuple $(i, E_{\leq t}, q, A, r)$: instance i; evidence available 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. The **unit of analysis is the resolved forecast**; but because forecasts are nested within instance and serially dependent, statistical inference is performed at the instance level (hierarchical models or instance-clustered resampling), and the number of independent *instances* — not the raw forecast count — bounds power for population-level claims. Forecast queries are issued by the organizer, not chosen by the entrant, so skill cannot be inflated by forecasting only the easy instances. Three assumptions are explicit: **A1** the future is partially predictable, bounded above [13,91]; **A2** latent target states are evaluated only through observable consequences; **A3** the instance changes, so the benchmark rewards adaptation. The measurement is agnostic to which latent structure produces the forecasts — the property that makes the grid of Section 6 architecture-neutral. A **forecast horizon** is part of the object: forecasts are specified at multiple horizons and abstraction levels — *short-horizon* concrete next states, *medium-horizon* routine deviations, and *long-horizon* abstract target-state transitions — reflecting that detailed prediction is reliable only at short horizons while long-horizon prediction requires abstraction. A hierarchical representation connects abstract long-horizon forecasts to concrete short-horizon ones, and each level is scored independently rather than assuming one fixed representation is optimal across horizons (Table 7). The four tracks (Section 8.2) form exactly this horizon–abstraction ladder, scored throughout against the routine baseline (R2) and the permutation specificity gate.

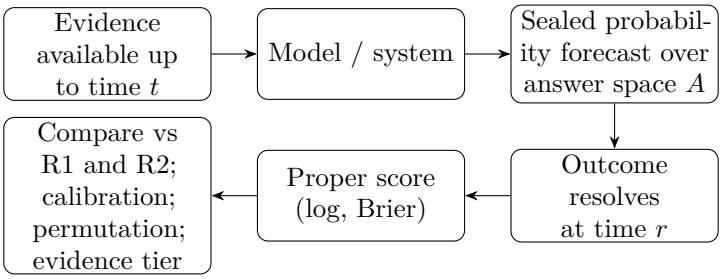

Figure 1: The TargetSpace evaluation loop. A model or system receives evidence available up to time t and emits a sealed (timestamped, SHA-256) probability forecast over the answer space A. After the outcome resolves at the resolution time r, the sealed forecast is graded with a proper score (log score in bits, Brier) and compared against the population-prior baseline (R1) and the own-routine baseline (R2), with the calibration gate, the permutation specificity gate, and the evidence-tier ablation. Only the sealed forecast and the resolved outcome are scored; the system's internal explanation or representation is not.

The capability as belief-state inference. It is useful to name what a competent system must internally maintain, while being precise that the benchmark never scores it directly. From observations $O_{1:t}$ of an instance, a system forms a **belief state** B_t — in the technical sense of a distribution over latent variables, updated as observations arrive and serving as a sufficient statistic of the history [88]. Here the latent variables are the target's evolving targets: active and competing goals and their priorities, binding constraints, salient recent episodes, stable preferences, recent target transitions, and the system's own uncertainty. The task is then: from $O_{1:t}$, maintain B_t and emit a calibrated distribution over the target-relevant future $O_{t+1:t+k}$ that B_t implies. **Crucially, B_t is latent and architecture-specific, and TargetSpace does not score it** — it scores

only the externally resolvable forecasts B_t produces, against sealed outcomes (A2; Section 5.2). The belief-state view thus motivates the object without reintroducing internal self-consistency as a criterion: two systems with very different internal B_t remain comparable because both emit a distribution over the same answer space A, and the own-routine baseline (R2) and the permutation gate test whether whatever B_t a system maintains is genuinely about *this* target rather than the population or the routine. A system may succeed with an explicit, inspectable B_t or with none that is human-readable; the benchmark is indifferent to the representation. B_t earns attention not because the benchmark inspects it, but because a forecast that beats R2 and collapses under permutation makes it a credible candidate object for later *explanatory audit* (Section 12) — a secondary analysis that never substitutes for prospective validation.

We keep several often-conflated notions distinct, because the benchmark scores some and not others. A **stated goal** is what a person declares they want; an **inferred goal** is what a model concludes they are pursuing; an **enacted priority** is what their allocation of a scarce resource (attention, time) actually reveals; a **predicted next action** is a surface behaviour; a **predicted future state** is a value in a defined answer space at a resolution time; and a **latent constraint** is a binding limit that shapes which transitions are possible. TargetSpace scores *predicted future states and their transitions* — with goal and constraint treated as latent structure evaluated only through those observable consequences (A2) — not stated goals, and not next-action mimicry. We use *goal-state transition* and the more neutral *meaningful transition* interchangeably where no theory of goals is required. Throughout, a **scored target** is a pre-registered discrete future state or transition with an externally observable **outcome** that resolves it; latent goals, constraints, and attention proxies are explanatory or evidential structure, scored only when mapped to such a target.

An **implicit target** is a latent orientation inferred from behaviour, context, repetition, avoidance, attention allocation, and transition patterns — a configuration the person behaves as if organizing around even when they have not declared, and may not have introspective access to [23], the corresponding goal. Implicit targets include unresolved concerns, avoidance patterns, preparation for a pending decision, social obligations, anticipated conflict, status preservation, uncertainty reduction, emerging commitments, and deferred or abandoned goals. Like stated and inferred goals, an implicit target is *evidence-bearing latent structure*, never a scored target in its own right: it earns credit only when mapped to a pre-registered future state with a deterministic, externally observable resolution (above), so "the person is avoiding this reply" is scored only as, e.g., $P(\text{no substantive response to message } m \text{ by } r)$.

5.2 The walk-forward sealed protocol

A retrospective benchmark replays logged histories, so apparent forecasting becomes partly retrieval — the failure that motivated contamination-resistant designs [16,17]. TargetSpace privileges a prospective, sealed mode (Figure 1): forecasts are timestamped and sealed (SHA-256) before outcomes exist and resolved automatically. We enforce strict walk-forward (prequential) evaluation — only evidence timestamped $\leq t$, indices rebuilt per slice; random cross-validation is prohibited. **Federation** is a core choice: the harness runs where the data lives, only sealed forecasts and resolved outcomes leave the client, reporting is aggregate-only — solving the dataset problem (most targets cannot be centralized) and the privacy problem at once. Two properties follow. First, target-specific forecasting is inherently **longitudinal rather than IID**: a target's earlier history may become relevant to later forecasts, so the evaluation preserves chronology rather than randomizing examples away from their temporal context. Second, a forecast is **scored against sealed external outcomes**, never for internal self-consistency: a system that produces confident rollouts inside its own latent simulator earns no credit until those rollouts resolve against what the

target actually did, which guards against a model that is competent only inside its own simulator.

5.3 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 forecasts the *evolving latent state that generates those outputs.* **Not event-outcome forecasting.** Sealed proper-scored forecasting of external events is established [28,40]; 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 goal 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 11): it evaluates such systems rather than replacing them.

5.4 Why target-state forecasting differs from generation and imitation

Four capabilities are easy to conflate, and the benchmark scores a specific one. **Generation** predicts surface outputs — the next token, frame, or fluent answer — judged by plausibility or fidelity. An **imitation or reactive policy** maps observations and instructions to actions, often from demonstrations or a learned policy, without necessarily exposing why one action beats another. **Consequence prediction** forecasts the future state that an action, event, or transition produces. **Planning** searches or optimizes over predicted futures to choose actions that reach a target state. These are not mutually exclusive — one system may generate, act, predict, and plan, and reactive and predictive components may coexist (a vision-language-action policy can be paired with a learned model that predicts subgoals [78]). TargetSpace directly scores *consequence prediction* and *target-state forecasting*: it does not infer planning ability from action success, and it does not require a model to expose a human-readable internal simulator. A system may participate if it can emit a calibrated distribution over externally resolvable future states — a reactive policy wrapped to report outcome distributions, a latent world model, an LLM, a symbolic planner, or a hybrid alike. This aligns TargetSpace with world-model evaluation at the measurement layer: it scores the forecast of the transition, not the reconstruction of a full future, the internal latent state, or the production of a plausible act.

The part of the future that matters. TargetSpace does not ask a system to predict every ripple in the river — every pixel, leaf, token, or surface detail, much of which is irrelevant or intrinsically unpredictable. It asks whether the system identifies and scores the *target-relevant* transition: the part of the future that matters for the specified target, horizon, and evidence tier. This is the same intuition that motivates predicting in a learned representation rather than reconstructing raw observations (Section 11.2), lifted to the measurement layer. It also fixes the benchmark's object precisely as a world-model question made architecture-neutral: given a **target**, **partial evidence**, and a **horizon** (optionally a hypothetical intervention), what distribution does the system assign to the target's **next state**? A system that can only produce a plausible surface, or reconstruct a full but mostly-irrelevant future, has not answered it.

An optional action-conditioned mode makes consequence prediction explicit. A base instance $(i, E_{\leq t}, q, A, r)$ (Section 5.1) extends to $(i, E_{\leq t}, q, A, U, r)$, where U is a candidate action, event, or intervention context, and the system predicts $P(z_{t+h} \mid E_{\leq t}, U)$, the target-state distribution at horizon h under U. Four things stay distinct: passive forecasting (no U), observational counterfactual prediction (a hypothetical U scored on later realized cases), actual intervention, and causal-effect

estimation. An action-conditioned forecast scored on observational data is *not* a causal estimate; causal identification requires intervention, natural experiments, or explicit structural assumptions (Section 12).

6 The Evaluation Grid: Evidence Tier × Architecture Class

Holding the forecast unit and scoring fixed, TargetSpace varies *what a system may observe* (evidence tier) and *what kind of system it is* (architecture class). Every cell is comparable because every system, whatever its internals, emits a distribution over the same A on the same sealed instance. Architectures that do not natively emit such a distribution — latent-predictive models, point-estimate policies, or human respondents — are scored through a fixed, disclosed **output adapter** that maps their native output to a distribution over A and is reported as part of the system, so the grid is architecture-*comparable under a disclosed adapter* rather than assumption-free. We did not invent architecture-neutral evaluation — representation-agnostic world-model evaluation was introduced by AutumnBench/WorldTest [41], and WorldPrediction [42] already scores latent-predictive and language models on shared instances. TargetSpace's addition to the grid is the target-specificity stack: calibrated, proper-scored, sealed evaluation with R1/R2 baselines and the permutation gate. Because systems differ in access and adaptation, each discloses whether it is frozen or adaptive, its retrieval, memory, and tool access, its prompt and adapter version, and its training cutoff where known; any adaptation may use only information available by the forecast cutoff, and leaderboards stratify systems by materially different access and adaptation regimes rather than implying a single unrestricted ranking.

6.1 Architecture classes

The grid below does not rank architecture classes in the abstract; it has nothing to say about which class is "better" in general. 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, a large language model, a vision–language model, a JEPA-style latent-predictive model, a symbolic or probabilistic system, a hybrid, and a human filling in a self-report are all evaluated at the level of the forecast output, not by inspecting internal representations.

Because the output adapter can change measured performance, it is disclosed and reported as part of the system under test. An output adapter maps a system's native output — an action, a token sequence, a latent prediction, a logical query result — to a probability distribution over A. The same underlying model paired with two different adapters is two different entries, and any tuning, temperature setting, or aggregation rule inside the adapter is part of what the score measures. We therefore require the adapter to be specified before the forecast is sealed, so the reported skill in bits over a reference reflects the system as actually run rather than a post hoc reinterpretation of its outputs.

6.2 The evidence ladder

The evidence ladder is the benchmark's measurement instrument (Figure 5). Tiers are cumulative; we give the canonical *person-domain* ladder and other domains substitute their own sensor ladders (Section 9), preserving the ordering from already-interpreted residue toward pre-interpretive observation.

Table 3: Architecture classes. LLMs and JEPA-style models are complementary components, not rivals: TargetSpace asks whether *any* of them produces calibrated, target-specific forecasts. 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**	**Forecast interface (native / via adapter / dependent)**
LLM-only	text-in, distribution-out language model	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	distribution via adapter; target-specific memory possible
JEPA-style / latent predictive (incl. world models)	learns latent dynamics, predicts in representation space [33,34,35,36,44]	state and action-conditioned prediction native; calibrated distribution via adapter; planning by search
Symbolic / probabilistic	explicit structured or Bayesian model	calibrated distribution and state prediction native; planning architecture-dependent
Reactive / imitation policy (incl. VLA)	maps observation + instruction to action [50,51]	action native; consequence distribution via adapter; not intrinsic
Hybrid	any combination (e.g. policy + learned subgoal model [78])	architecture-dependent; reactive and predictive components may coexist
Human self-report *(baseline)*	the target, or an expert, predicts itself	distribution via elicitation; not ranked
Oracle upper bound *(baseline)*	a privileged predictor; approximate ceiling	not ranked

The ladder also encodes a distinction richness alone does not. Lower tiers (L0–L1) are records the target already produced by deciding what to write, send, or schedule — they begin *after* relevance was assigned, and largely encode routine. Higher passive tiers are **pre-interpretive**: captured before the target fixed what would matter, they carry **retrospective option value** — the same recording can be re-scored under questions chosen later. Pre-interpretive capture is not thereby 'objective'; it is, more precisely, *independent from retrospective human interpretation.* The target-specific hypothesis is sharp: **digital exhaust may predict routine; passive evidence may reveal deviations from routine** — exactly the transitions where skill over R2 is earned. 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.

Departures from an individual's established routine are not, by default, noise. Research in neuroscience and complex-systems biology indicates that within-system variability can carry information about latent state and cognitive condition [67], and arises from multiple sources that need not be mere nuisance variance — some of it functional, with possible roles in adaptation or exploration [68]. Distinct internal configurations can also produce similar observable outputs, a property known as **degeneracy** [69] — the capacity of structurally different elements or mechanisms to perform

Table 4: The person-domain evidence ladder. By reporting target-specific skill (over R2) as a function of tier, TargetSpace turns 'does richer observation add target-specific information?' into a measured quantity rather than an argument.

Tier	Evidence (person domain)	Sensitivity
L0	calendar + communications metadata, routine, digital exhaust	social graph / rhythms
L1	text traces (chat, notes, documents)	high — third-party content
L2	audio / transcripts	very high — bystanders
L3	passive multimodal observation (egocentric / scene video, ambient audio, screen)	extreme
L4	location / behavioral / mobility traces	extreme — re-identifying
L5	physiological / specialized sensors	extreme — health-revealing

the same function or yield the same output (not the model collapse, numerical degeneracy, or redundancy the term denotes in machine learning). In neural systems this is concrete: closely matched network activity can arise from substantially different underlying circuit parameters [70]. (These results concern neural and biological systems; we extend them only by analogy to behavioural forecasting.) Two targets may therefore reach the same state along materially different trajectories, so population averaging or endpoint-only evaluation can erase precisely the person-specific structure that signals an emerging target-state transition. TargetSpace accordingly *preserves* a departure from an individual baseline so the evidence remains available for analysis; this never alters a sealed forecast or its scored target, both fixed at the forecast time (Section 5.2). Whether a departure mattered is asked of preserved evidence in separate, clearly-labelled exploratory analysis, not a post-hoc relicensing of the primary score.

The ladder is more than a convenience stratification; it reflects different partial, mediated windows on the target. Textual records are *translated* traces — experience already rendered into language; logs and digital exhaust are *behavioural* traces; audio and video are *temporal observational* traces; and action-conditioned or interventional evidence exposes how the target responds when the world pushes back. No band is reality, and a higher tier is not automatically better — whether it pays is the empirical question the ablation answers. Because the relevance of an observation may only become clear retrospectively, TargetSpace treats **raw-evidence preservation, provenance, and ablation as first-class evaluation concerns**: the benchmark can ask not only whether a system used evidence but *which* evidence was necessary to improve a target-specific forecast. Preservation is always bounded by the consent, federation, aggregate-only, and retention limits of Section 12; the benchmark scores evidence value, it does not license indiscriminate retention.

7 Metrics

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 1.2 — a distribution concentrating — with no physical meaning.

7.1 Preserved foundation (adopted)

Forecasts are scored with strictly proper rules: the **log score** (primary) and **Brier score** (secondary) [14,15,18]. The headline quantity is **Skill**, in bits:

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

a paired comparison on identical sealed instances, equal in expectation to an information gain — equivalently, the prospective predictive code-length gain of the system over the reference — **not a novel metric**. It is reported against **R1** (population prior; the entry condition) and **R2** (the target's recency-weighted, walk-forward, instance- and task-specific routine; the target-specific condition). R1 uses population information only, estimated without the scored target's own history (leave-one-out), matched on question type, answer space, and horizon, rolling, with smoothing for rare classes; its exact fitting is pre-registered. Forecasts 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. **Calibration** gates the result, in the spirit of holistic evaluation that treats it as first-class [43]: it is assessed by a pre-registered combination of calibration intercept and slope, a reliability (proper-score) decomposition, and an uncertainty-aware summary, with top-label ECE bands ($\leq$0.10 pass / $\leq$0.20 warn / $>$0.20 fail) as preliminary diagnostics; any recalibration uses a separate pre-seal split, is disclosed, and is never fitted on sealed outcomes. The **permutation specificity gate** scores a system's model for i against another instance's outcomes — matched and stratified within compatible question types, horizons, and answer spaces, swapping complete eligible histories rather than scrambling outcomes, so target identity is not confounded with differing base rates; skill that does not collapse was not target-specific. R2 is a **pre-registered**, organizer-fit (not entrant-fit), frozen, walk-forward model of the target's own routine (default v1 recipe in Appendix B) — marginal frequencies, persistence, recent transitions, time-of-day and day-of-week effects, and recurring calendar/deadline commitments known before the forecast — fixed before scored evaluation and never selected on test results; it is intended as the strongest *simple* routine model, not an unrestricted learned personal model, and is admitted only where it itself beats R1, so 'skill over R2' cannot be manufactured against a weak baseline (its exact feature set, recency kernel, minimum history, and admission rule are pre-registered). The permutation gate is reported as an effect — true-instance Skill minus the mean permuted-instance Skill, as both absolute and proportional skill loss with an interval and a pre-registered margin — whose resolving power grows with the number of instances permuted, so it is directional and operational in the five-person pilot and a powered test only above a stated minimum cohort size. Uncertainty respects the repeated, serially dependent structure of forecasts 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 forecast or day — is the independent unit for between-person claims; the current reference implementation reports point estimates. Of this stack, the **R2 own-routine baseline and the permutation gate** are the unoccupied elements; R1, proper scoring, calibration, and sealing are adopted from forecasting practice [14,15,18,28,40,43]. A deliberate consequence of scoring a distribution over a defined answer space is that **exact surface form is not the success criterion**: a system is judged on proper score, calibration, rank/order, and transition correctness, not on reproducing a reference phrasing. This mirrors the observation from the JEPA literature (Section 11.2) that predicting the *embedding* of a target answer rather than its exact wording avoids penalizing answers that are correct but phrased differently; we adopt the principle at the level of scoring.

7.2 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) [29] 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 5: 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 [15]
Log + Brier, ECE gate, sealing	adopted	[14,18,28]; calibration-as-axis [43]
R2 own-routine + permutation gate	**anchor (this work)**	own-history baselines & self-consistency [5] as cousins
Top-k recall; transition-path likelihood	adapted	retrieval; cell-fate / trajectory likelihoods [30]
Time-to-correct-resolution	operationalization, not new idea	goal-recognition design / WCD [29]; online goal recognition
False-resolution rate	new metric (concept has precedent)	premature-commitment / false-stop / FPR in goal recognition

7.3 A worked instance (synthetic, illustrative)

To make the apparatus concrete — with invented numbers, no participant, and no result claimed — Figure 2 works through one instance. 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 forecast 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.)

8 First Instantiation: Personal World Modeling

Personal world modeling is the instantiation we carry furthest, because understanding individual people is valuable and the observation bottleneck bites hardest there. Here the target is most vividly a lived trajectory, and the passive evidence tiers (Table 4) are its observable traces — attention, context, interruptions, pressure, and the timing of action — from which a model must infer the latent target states and forecast their transitions, never accessing the person's experience itself. The target-system instance is a consenting individual; the scored target states are categories such as

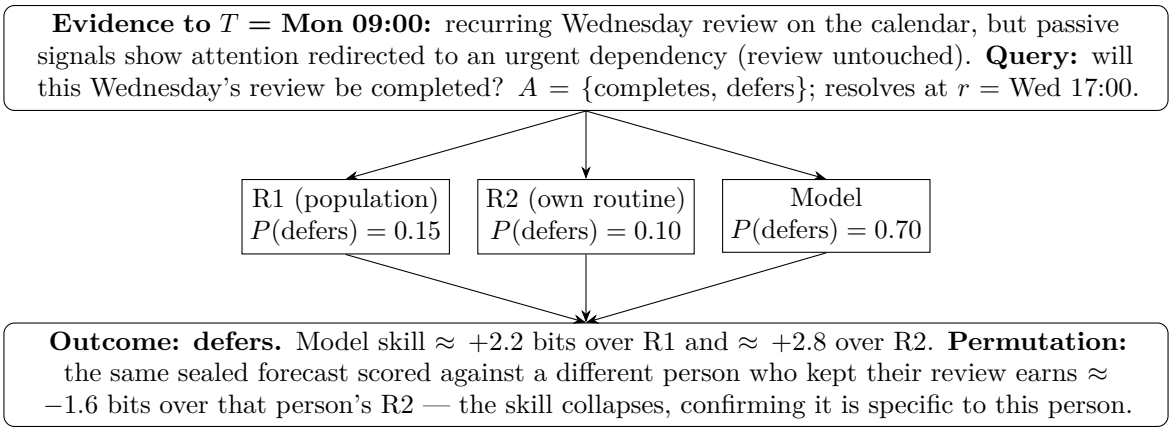

Figure 2: Worked synthetic instance of a target-state transition forecast (the same instance described in the text). At sealed time T (Monday 09:00) the evidence shows a recurring Wednesday review on the calendar but passive signals indicating redirected attention. The query asks whether the person completes this Wednesday's review (answer space A = {completes, defers}), resolving at r (Wednesday 17:00). Three forecasts assign P(defers): R1 (0.15), R2 (0.10), and the evaluated model (0.70), which reads the attention shift. The outcome is *defers*, giving the model about +2.2 bits over R1 and +2.8 bits over R2. A permutation check scores the same sealed forecast against a different person who kept their review, yielding about −1.6 bits over that person's R2: the skill collapses, confirming it is specific to this person. All numbers are illustrative and synthetic.

commitment active versus abandoned, priority maintained versus displaced, and task continuation versus switch (Table 6), while attention, concerns, and avoidance are *evidence* about those states rather than states themselves (the evidence ladder is Table 4). The personal-AI question is sharpened by target-specificity: **not merely 'does passive capture help?' but 'does passive first-person evidence help a model forecast *this person's* future target-state transitions beyond population priors and beyond this person's own routine?'**

Concretely, the **forecast 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 6); attention allocation, social context, and emotional cues are **evidence**, and deadlines, dependencies, and competing obligations are **constraints**. The target-specific hypothesis is the same throughout: digital exhaust (L0–L1) often overrepresents this person's routine, which a strong R2 already captures; passive evidence (L2–L3) may reveal *deviations* from routine — the shifted attention, the commitment about to slip — that are exactly where skill over R2 is earned.

The generation / imitation / forecasting distinction (Section 5.4) maps cleanly onto the personal case. The task is not to *summarize* a person (generation) or to predict *what people usually do* (a base rate, or the imitation of typical behaviour); it is to forecast *this* person's deviations and next-state transitions from partial evidence, while beating that person's own routine baseline.

8.1 Attention as evidence and as process

Attention has two roles here, and the benchmark must test rather than assume the second. **As evidence**, attention allocation reveals what currently competes for a person's limited cognitive and behavioural resources [24]: how someone spends a scarce resource is a revealed signal of what

they are trying to do, and its trajectory reveals it continuously, often before any action resolves — a leading indicator of target-state transitions. Because the same allocation can serve many goals, attention is evidence of allocation, not a read-out of goal identity or ground truth. **As process**, selective attention changes which information is sampled, represented, rehearsed, learned, and turned into action [71,72], and through those effects it can stabilize, revise, displace, or help constitute the next target state (Section 1.3); goals direct attention, and attention reciprocally maintains and updates goals [73]. We use *participatory* only in this bounded sense — attention *participates in* and *reshapes the conditions under which* target states form; it does not create them from nothing. (In non-person domains the analogue is whatever scarce resource the instance allocates and how that allocation is structured — compute for a software agent, capital or managerial attention for an organization [74].)

Attention-conditioned predictive lift. Because the process role is an empirical claim, we make it falsifiable rather than assume it. Writing the transition generically as $P(z_{t+1} \mid z_t, a_{\leq t}, e_{\leq t}, u_{\leq t}, c_{\leq t})$ — with a attention, e evidence, u action, and c constraints — we compare a matched reduced model $P(z_{t+1} \mid z_t, e_{\leq t}, u_{\leq t}, c_{\leq t})$ that omits the attention trajectory. **Attention-conditioned predictive lift** is the difference in proper-scored, target-specific Skill between the two, measured beyond R1, R2, stated goals, and contemporaneous behaviour. The protocol fixes what counts as an attention proxy and at what temporal granularity, how missingness is handled, how the proxy is separated from mere activity frequency, and how leakage is prevented; lift is scored only on sealed future outcomes. A positive lift demonstrates incremental predictive value, *not* causation: in an observational design attention may proxy an unobserved cause, so causal identification would require intervention or a natural experiment (Section 12). The active-inference account, in which action and attention (as precision) shape expected future states, is one compatible framework [19,32,75], not a settled mechanism.

8.2 The corpus bottleneck and the capture-to-labels stack

The missing data layer named in Section 1.2 is concrete in this domain. Digital exhaust (L0–L1) records a person's *outputs* — what they wrote, sent, or scheduled — not the process that produced them. Passive first-person observation (L2–L3) is the stream from which goals, constraints, interruptions, and transitions could be inferred *before* intent is formalized: passive audio, a screen / digital-activity timeline, documents and files touched, calendar and task context, interruptions, spoken intent, environmental and social context, and the timing and sequencing of actions. Egocentric corpora such as Ego4D and Ego-Exo4D [56,57] show passive first-person capture at scale, but as generic clips for activity recognition — not longitudinal records of one tracked individual aligned to that person's goals and outcomes over time.

Raw capture is not, by itself, a benchmark. Turning observation into evaluable state-transition data takes a stack — passive sensing, a longitudinal multimodal corpus, segmentation with goal/task/outcome/transition labels, a temporal state representation, the prediction benchmark, and a privacy-filtering layer (Section 12). **TargetSpace is the evaluation layer**: it fixes the evaluable target and the scoring spine and specifies what the upstream layers must deliver, but does not supply the corpus. Assembling a labelled, privacy-filtered, goal-aligned longitudinal corpus is the open bottleneck, and the pilot (Section 8.3) builds only a thin slice; passive recording alone is insufficient. For high-sensitivity tiers (L2–L5), federation is necessary but insufficient: a viable implementation would likely require on-device filtering that converts raw audio, video, and screen streams into task-relevant semantic events before persistence — mitigating, though not solving, the bystander-consent and recording-law problems of Section 12. **Hardware is the capture layer, not the benchmark**: wearable and passive sensors raise stream fidelity, but TargetSpace takes

Table 6: A taxonomy of target-state transitions the personal track forecasts. 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 forecast range.

Transition type	**Operational description / resolution criterion**	**Label-ambiguity risk; evidence needed**	**Likely 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/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	attention-conditioned; medium
Competition	two targets contend; resolved by which receives sustained allocation/follow-through	high; attention + outcomes	attention-conditioned; short–medium
Displacement / priority reversal	a new target supplants the prior one; resolved by a switch in sustained allocation	medium; attention + calendar/task	attention-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/opportunity creates a target; resolved by uptake after the encounter	high; evidence event + uptake	R1/R2 weak; short–medium

no position on a device; whether richer capture pays is the evidence-tier ablation (Section 6.2). A human-centered world model must be learned from observation, since language and digital exhaust alone are too sparse and post-hoc.

8.3 Tasks and a planned validation study

Version 1 uses auto-scorable, high-frequency targets organized into four domain-neutral tracks 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. This paper reports no results. The first instantiation, an audio-first personal pilot, is an **exploratory feasibility and variance-estimation** study, not a confirmatory test of the cross-domain or cross-architecture thesis: five consenting participants, thirty days, sealed daily calibrated predictions, systems differing only in evidence (A population prior; B digital exhaust; C +text/transcript; D +continuous first-person stream; model and prompts fixed across B–D). Its purpose is to check that the apparatus functions — forecast generation and sealing, reproducible outcome resolution, sufficient instance volume, and the variance and dependence structure of score differences — and to supply the inputs for *simulation-based* power planning of a later confirmatory study; with five participants it is not powered to confirm population-level effects,

establish generalization, or estimate subgroup or cross-domain claims. A strong own-routine baseline compounds this: because a near-deterministic R2 leaves little headroom, detecting reliable lift over it may require high-frequency longitudinal observation and larger cohorts than the pilot provides. *Research hypotheses for the program* (tested confirmatorily only in a later powered study, and treated here as *exploratory measurements*): B–D beat R1 and R2 (H1); skill is non-decreasing in evidence, the open question being the C→D increment (H2); skill collapses under permutation (H3); null: no system beats R2 (H0). The primary endpoint is the paired prospective log-score improvement over R2, $\Delta L(M, \mathrm{R2}) = L(\mathrm{R2}) - L(M)$ in bits per forecast, with within-person day-blocked and between-person person-clustered uncertainty; performance over R1, calibration, and target-specific skill loss under matched permutation are validity gates. The attention-allocation family (TS-A) is the primary task family and commitment follow-through (TS-D) a key secondary; the final primary family and multiplicity rule are fixed in the pre-registered protocol. **Abandonment criteria:** no system beats R1 → noise; none beats R2 → target-specific forecasting not demonstrated; skill survives permutation → not target-specific.

Pre-registration and controls. Forecast instances are organizer-issued under a pre-registered eligibility and sampling frame — not entrant- or retrospectively selected — stratified across routine-continuation and transition cases and de-duplicated where evidence or outcome windows overlap; skill is reported separately on routine versus transition instances, since skill concentrated in transitions is stronger evidence of structure beyond routine replay. Outcomes are resolved by pre-registered rules: deterministic behavioural outcomes are preferred, ambiguous transition classes use forecast-blind adjudication with reported inter-rater reliability, unresolved cases follow a rule fixed before forecasts are examined, and an outcome label is never the same self-report channel a model uses as evidence. Missing evidence and unresolved outcomes may themselves be informative; their handling and sensitivity analyses are pre-registered, missingness is usable as evidence only when observable before the cutoff, entrants may not drop organizer-issued instances, and abstention is scored as a reference distribution under the same rule for human and machine systems. Negative controls comprise target permutation, a leakage / known-null canary, and a temporal-specificity (wrong-time-window) check. Before any data collection the operational protocol pre-registers the R1 and R2 fitting, sampling frame, answer spaces, outcome-resolution rules, calibration split, permutation matching, missingness handling, abstention rule, primary endpoint, multiplicity rule, negative controls, and statistical analysis; full across-context, across-regime, and across-population transfer is later work, not a v1 gate.

Table 7: A horizon–abstraction hierarchy for target-state forecasting. 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

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 forecast should not contradict the abstract target state without explicit uncertainty or branching); *refinement* (as the horizon shortens and evidence arrives, abstract forecasts 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 forecast); and *rollout drift* (repeated one-step predictions may diverge from external reality even when each step is locally accurate [76]).

8.4 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 [33,36]. 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 goal state, hidden constraints, the likely next transition, and the next useful action? It is a world model whose dynamics are a human task process, scored by calibrated skill over R1/R2 rather than by reward or reconstruction.

A behavioral world model, not a physical one. The world-model idea (above) earns its power because predicting the consequences of actions lets an agent test a plan before acting [12], with planning as search over predicted futures [37,38]. TargetSpace generalizes this predictive substrate, not its physical content: it asks for no prediction of torques, contacts, or pixel dynamics and claims no understanding of a physical world it never observes. The classical form is *state + action → next state*; the TargetSpace form is *passive longitudinal observation → belief state over latent targets* (Section 5.1) *→ target-relevant future behaviour*. A competent system must therefore learn a **compact behavioral world model** of one observed person — equivalently a personal, or target-state belief, model — capturing recurring contexts, constraints, stable preferences, salient episodes, unresolved decisions, and target transitions, and predicting how that target state evolves as evidence arrives. The analogy is to the *form* of the predictive state, not its physical realization; like any learned dynamics model it can compound error over long horizons [76], and only its sealed forecast is scored — not planning, action, or the model itself (Sections 5.4, 12).

The shift is clearest by contrast with what existing evaluations ask. A **transcript** or chat benchmark asks *what was said?*; a **memory** benchmark asks *what happened?*; TargetSpace asks a third question — *what goal state is the person in, what transition is likely next, and what intervention would help or harm?*

Retrieval and long context are not the capability. Two now-standard memory paradigms make the same point concrete. **Retrieval-augmented generation** conditions a model on documents fetched from an external store [90]; **long-context** models simply hold more of the history in the window. Both surface *what happened* more readily — usefully — but neither is scored on whether the system inferred *why* an episode mattered or *what target* it implies: a retrieval system can surface the reread email, whereas TargetSpace asks whether a model reads it as an avoided decision and forecasts the slip. That is why a higher recall@k or a longer context window does not by itself move prospective skill over R1/R2 under the permutation gate; the architecture grid (Section 6.1) accordingly places base, long-context, retrieval, episodic-memory, and belief-state–updating systems on identical sealed instances.

Prompts may reveal goals but often omit the **pre-prompt** process — a prompt is a late, compressed artifact emitted *after* an intention has partly formed — which passive observation can preserve (attention, interruptions, partial actions), and the evidence-tier ablation tests whether

preserving it improves prediction. TargetSpace does not build this user-centered world model; it operationalizes it as evaluation: fixing the latent quantities to predict, the baselines to beat (R1/R2), and the calibration and permutation gates that separate a genuine model from a fluent guess. The target is abstract predictive state, not raw reconstruction, and the framing is architecture-neutral (Section 6.1, Table 10).

8.5 Asymmetric observability: self-report, observation, and person knowledge

Personal intelligence forces a question the benchmark must handle carefully: who models the person better — the person, or an outside observer? We reject any claim of observer superiority. An individual has privileged access to immediate subjective experience, private beliefs, felt emotions, conscious intentions, and autobiographical meaning. A longitudinal observer can sometimes detect repeated behavioural regularities, inconsistencies between stated and enacted priorities, avoidance patterns, cross-situational structure, and slow changes that are hard to notice from inside the stream of experience. Self-report and observation are therefore **partially overlapping but non-redundant evidence channels**. **The issue is not observer superiority, but asymmetric observability.** Their relative predictive value is an empirical question the benchmark should measure, not assume.

This mirrors person-perception research. The self–other knowledge asymmetry model holds that self and informants know different things — the self more accurate for internal, low-observability traits, informants for evaluative, highly observable ones [58] — and meta-analyses find informant reports valid and often incremental to self-report, for some outcomes exceeding it [59,60]. Self-report has limits of its own: people often lack introspective access to what drives their behaviour [23], which motivates real-time experience sampling [61] and passive 'personal sensing' [62,64]; internal and external self-awareness are themselves distinct constructs [63]. None of this makes observation more accurate than self-report — it makes the two complementary.

TargetSpace makes the comparison *prospective and empirical* rather than rhetorical. Self-report is an evidence modality and an architecture class (human self-report), and a baseline — not a gold standard; passive observation is another modality, not automatically more accurate. The evidence-tier ablation, and the head-to-head of human self-report against observational systems on identical sealed instances, measure which modality — or which combination — yields more calibrated, person-specific skill for a given target. We expect a structure rather than a winner: **first-person access may be richer locally, while longitudinal external observation may be more informative structurally for some prediction targets**, and combinations of modalities may outperform any single one. The benchmark is built to find out, not to presume; clinical observation is the right analogy — complementary evidence, not clinician omniscience.

8.6 Implicit Targets and Self-Narrative

Surface action is an ambiguous signal of what a person is organizing around, so it helps to separate three layers. The **action layer** is what the person does (rereads a message; reopens a document; revisits a calendar event). The **narrative layer** is how they frame it ("I'm just making sure I understand it"). The **target layer** is what the behaviour is actually organizing around — which may be avoiding a difficult reply, reducing uncertainty before a decision, preserving a relationship, or protecting status. TargetSpace is primarily a benchmark for the third layer, using the first two as evidence. The separation is not decorative: declared intentions predict behaviour only weakly (the intention–behaviour gap [84]), motives inferred from behaviour diverge systematically from self-attributed ones [85], people may not introspect the causes of their own behaviour (Section 8.5),

and self-narration — inner speech that mediates attention, monitoring, and self-regulation [89] — can report a goal other than the one the longitudinal pattern reveals. A model that predicts only the action layer, or that trusts the narrative layer at face value, will miss the target layer exactly where it matters.

This is where surface behaviour becomes informative only longitudinally. Considered once, rereading an email, drafting but not sending a reply, repeatedly checking a thread without responding, or shifting attention as a deadline nears are each individually ambiguous. Across a tracked history they can mark a latent target under tension — an unresolved decision, an avoidance pattern, a commitment about to slip — precisely the deviations from routine where skill over R2 is earned. The benchmark does not ask a model to *narrate* this latent state; it asks it to convert these traces into a calibrated forecast of a pre-registered observable outcome (Section 5.1) and certifies, through R2 and the permutation gate, that the resulting skill is specific to this person rather than to population-level patterns of attention or delay. The cognitive-science framing here is motivation, not mechanism: TargetSpace takes no position on theories of inner speech or self-narrative and stands if they are revised — it requires only that some behaviourally relevant targets are implicit and inferable from passive longitudinal observation, which the evidence-tier ablation then tests rather than assumes.

9 Tracks: A Multi-Track Apparatus

TargetSpace is not a single dataset and not one domain-specific benchmark. It is a shared evaluation apparatus with multiple application tracks. Each track defines domain-specific *targets, evidence sources, horizons, baselines, and outcome validators*, while preserving a common scoring spine. This mirrors how disciplined evaluations are organized — a fixed metric spine applied across scenarios (as in holistic LLM evaluation [43]) or across domain areas (as in systems benchmarking) — rather than as one monolithic task.

Naming. TargetSpace is the framework; **personal intelligence** is the flagship capability and track (the TS-Personal track); the remaining tracks show the framework is not merely a personalized-assistant benchmark.

The shared scoring spine is invariant across every track: a tracked target object; a partial evidence bundle; a forecast horizon; a sealed, walk-forward prediction; proper scoring; a calibration gate; an own-routine/own-system baseline (R2) and a population baseline (R1); a permutation specificity test; an evidence-ablation ladder; and provenance with deterministic outcome validation (Sections 3, 5, 7). A track 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. We cap the top level at a small number to avoid the dilution that afflicts many-track benchmarks, and we tier each track honestly by readiness.

Adding a track instantiates the spine, not changes it. We **deliberately exclude** scientific perturbation forecasting (e.g. cell-fate): it is one-shot and cross-sectional, its strong baseline is a cross-sectional mean rather than an own-routine R2, and its target is not tracked over time, so it fails the spine. Energy and markets are one track with two regimes, not two.

10 How to Use TargetSpace

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 forecast 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 forecast times and horizons;

Table 8: The five TargetSpace tracks, one shared spine. *Status* — **current**: apparatus exists (synthetic, pre-pilot) for the person track only; **planned**: strong sealed precedent and a natural own-routine baseline, not yet implemented (energy/grid: GEFCom [52]; physiology: PhysioNet/CinC [53], OhioT1DM/BGLP [54]; personal: GLOBEM [55]); **research**: a distinct regime exists but a proper-scored forecasting protocol and/or a strong R2 are not yet established. We make **no claim that any track other than the synthetic person track is implemented**, and no claim that TargetSpace solves robotics, healthcare, energy, or enterprise forecasting.

Track ***(status)***	**Target object**	**Example evidence bands**	**Horizon**	**Example target states**	**Validator / outcome**
TS-Personal *(current)*	a consenting individual	metadata, text, audio, passive multimodal, location, physiology	hours–weeks	commitment status, priority shift, routine deviation, task switch	observed action / confirmation
TS-Health *(planned)*	a patient	vitals, labs, continuous glucose monitoring (CGM), wearables, notes	minutes–days	glycemic excursion, deterioration onset, care-state transition	clinical onset labels / sensor thresholds
TS-Energy *(planned)*	a series / asset	history, weather, calendar, exogenous covariates	hours–days	load / renewable level band, price regime	realized value / market settlement
TS-Robotics *(research)*	embodied agent + scene	proprioception, sensor stream, action log	sub-second–minutes	goal-conditioned configuration, subgoal transition	achieved configuration
TS-Enterprise *(research)*	a project / team / workflow	trackers, commits, comms metadata, releases	days–quarters	milestone state, scope / priority drift	observed milestone / outcome

(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 forecasts**, the **horizon**, and the **target domain**. 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 understanding from generic prediction at a glance.

11 Related Work

11.1 What prior work already owns (and we adopt, not claim)

Credibility requires being explicit about what TargetSpace does *not* invent. We adopt and cite, rather than claim, each of the following.

11.2 Memory, belief states, and goal inference

Three literatures bear on what TargetSpace asks a system to maintain, and clarify what it does *not* reward. **Memory.** Cognitive science distinguishes episodic from semantic memory [83] and, more pointedly here, treats autobiographical memory as a goal-driven *construction* rather than a stored replay [81], with the same constructive system used to simulate the future as to reconstruct the past [82]. This is why we frame the target capability as forward-looking belief maintenance rather than archival recall (Section 1.2). **Belief states.** Maintaining a distribution over latent variables,

Table 9: What TargetSpace adopts rather than claims. Its remaining contribution is target-specific forecasting under strong controls — the R2 own-routine baseline and permutation specificity gate — assembled around one evaluation question.

Ingredient	**Owned / established by**	**TargetSpace stance**
Representation-agnostic / architecture-neutral world-model evaluation	AutumnBench / WorldTest [41]	adopt the framing; add calibration + R1/R2 + permutation
Cross-architecture comparison on shared instances	WorldPrediction [42] (ran JEPA + LLMs)	adopt; add proper scoring + sealing + specificity
Latent predictive representation learning	CPC [44]; JEPA [33–36]	precedent for the latent-predictive class; we evaluate it
Proper scoring & calibration as a measured axis	Good/Brier/Gneiting&Raftery [14,15,18]; HELM [43]	adopt directly
Sealed / prospective / leakage-free forecasting	ForecastBench [28]; Prophet Arena [40]; SWE-bench [2]	adopt
Evidence ablation (e.g. active vs passive sensing)	digital-phenotyping forecasting literature	adopt; report as lift over R2
Goal recognition & time-to-identification / resolution timing	goal-recognition design / WCD [29]; online goal recognition	adopt; re-cast as proper-scored, prospective, on tracked instances

updated from partial observations and serving as a sufficient statistic of the history, is the belief-state formalism of partially observable decision processes [88]; TargetSpace adopts the formalism as motivation (Section 5.1) but scores only the externally resolved forecasts a belief state produces.

Latent-goal inference. That observers recover latent goals and beliefs by inverting a model of approximately rational action is the Bayesian theory-of-mind / inverse-planning account [86,87], the cognitive-science complement to the machine theory-of-mind [6] and goal-/plan-recognition [29] work discussed above; behaviour is further organized by implicit motives that diverge from declared ones [85] and by intentions that predict action only weakly [84]. TargetSpace differs from all three in object and protocol: it scores neither recall (memory benchmarks), nor the internal belief state (world-model evaluation), nor which goal from a fixed set best explains past behaviour (goal recognition), but the prospective, calibrated, instance-specific forecast of the next target-state transition, under R1/R2 and the permutation gate.

11.3 World models, predictive learning, and consequence forecasting

A long lineage predicts the *latent* state rather than the surface: predictive coding [19,39]; world-model agents that plan over learned latent dynamics [12,37,38], with hierarchical variants planning over abstract latent subgoals [77]; joint-embedding and contrastive self-supervised methods [44,45,49] and their information-theoretic and recurrent-world-model precursors [47,48]; LeCun's JEPA program [33–36], a *training framework* that predicts a target's *embedding* (optionally action-conditioned [36,46]) rather than reconstructing pixels or tokens, now extended to language targets [79]; and vision-language-action policies, in which broad, dexterous action emerges from multimodal policy learning [50,51], increasingly coupled to generative subgoal models [78]. The motivation bears on our object: predicting in embedding space keeps salient structure and discards the unpredictable — the same reason TargetSpace scores *transitions in a latent target state* rather than full-future reconstruction.

Reactive, predictive, planning, and hybrid systems are all candidate participants, not competitors, and we take no side on which will prevail. Uniformly, though, they are compared by representation

quality, reward, task success, action accuracy, or rollout quality — evaluations that do not, by themselves, test calibration, skill over a target-specific routine, instance specificity, prospective sealing, evidence-value attribution, or multi-horizon target-state forecasting. TargetSpace supplies that measurement layer and applies it to these classes on identical sealed forecast instances; what matters for scoring is whether the predicted transition resolves correctly against the external target, not how the prediction is represented internally.

Table 10: Positioning, not a ranking. The three families are complementary; TargetSpace does not solve world models and takes no position on the eventual architecture. It is architecture-neutral (Section 6.1) and asks whether passive observation improves calibrated prediction of human goal-state transitions.

Dimension	**Conventional LLM benchmark**	**Robotics / world-model benchmark**	**TargetSpace**
Input	a prompt / fixed context	state + action (sim or embodied rollout)	passive observation of a tracked instance up to a sealed time T
Prediction target	next token / answer	next state; consequence of an action (often in latent space)	latent target state and its transition
Evaluation object	output correctness / quality	task success; rollout / planning accuracy	calibrated skill over R1/R2 + permutation specificity
Time horizon	single turn / short	short-to-medium rollouts	hours–weeks; walk-forward, sealed
Role of language	central (it is the task)	peripheral (instructions)	one evidence stream among many; not the success criterion
Role of passive observation	none	sensor stream for control	central — the substrate; its value is measured by the evidence-tier ablation
Failure mode	fluent but wrong; contamination	reconstructs detail; rollout drift; sim-to-real gap	predicts the average not the target (fails R2 / survives permutation)
Success criterion	matches reference / human preference	reaches goal state; low rollout error	calibrated skill over R2 + permutation pass (beats routine, stays calibrated, is target-specific)

11.4 Forecasting, person modeling, and goal recognition

Calibrated event forecasting is established — ForecastBench [28], Prophet Arena [40], and others — 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: Park et al. [4,5] replicate individuals' survey answers (their own-consistency reference is the closest cousin to R2, though used as a ceiling, not a baseline); KnowMe-Bench [26] formalizes person understanding but as *retrospective* comprehension, not prospective forecasting; EgoToM [27] infers a camera-wearer's goals and future actions from egocentric video, but per clip, by accuracy, without calibration or a persistent target; PersonaMem [7] tracks evolving preferences. Related person-modeling and digital-twin efforts — personalization [3], preference modeling [8], digital twins of survey and behavioural respondents [9,10], a foundation model of human cognition [11], and multimodal theory-of-mind question answering [31] — predict held-out responses or preferences rather than sealed future target-state transitions of a tracked instance. Person-specific multi-horizon forecasting with active/passive evidence comparison already exists in digital phenotyping — we concede this and differentiate on proper scoring, calibration, and the permutation gate. Goal- and

plan-recognition [29] infer latent target/goal posteriors under partial observability and define how early a target is identifiable; the field also studies online recognition and changing, hierarchical, or partially observable goals, so we do *not* claim prior work assumes goals are fixed. The narrower distinction is in the evaluation object: 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 interaction with the environment produce the *next* target-state transition — emergence, stabilization, competition, displacement, abandonment, resumption, or opportunity-induced creation (Table 6) — 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; the extension from recognizing a current goal to forecasting target-state formation is subordinate to the target-specificity stack, not a separate theory of goals.

11.5 Cross-scale target-state systems

Beyond persons, the target-state idea recurs — active inference and preferred states [19,32], cell-fate prediction [30], and Levin's morphogenesis [25] (Section 4) — which we cite as lineage and sibling domains, not as competing calibrated benchmarks.

11.6 Intelligence measurement and ARC

Our framing is indebted to Chollet's account of intelligence as skill-acquisition efficiency and to the ARC-AGI line's contamination-resistant, private-evaluation design [65,66]; TargetSpace shares the commitments to novelty, freshness, and efficiency (the evidence-tier ablation asks how much evidence a system needs, not only whether it eventually succeeds). ARC evaluates adaptation within a given problem frame, whereas TargetSpace asks the earlier question of inferring the latent state and target from observation when the goal is not framed; both are needed, and neither subsumes the other. Because static, fixed-answer benchmarks saturate and can be gamed once their targets become public [16,17,28], the prospective, future-resolved measurement regime is itself part of the contribution, not merely hygiene: target-specificity is scored against outcomes that do not exist at forecast time.

11.7 The conjunction

Each ingredient above is occupied somewhere; the conjunction is not. We state the supported claim narrowly: **to our knowledge no existing benchmark scores prospective, calibrated, proper-scored forecasts 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.** We do *not* claim to be first at architecture-neutral evaluation [41], cross-architecture comparison [42], calibrated sealed forecasting [28], evidence ablation, or resolution-timing [29]. The contribution is their assembly around one measurable question.

Table 11 audits neighboring benchmarks against the five dimensions. The labels are cautious design judgments, not numerical scores: *clear* (a primary, explicit target), *partial*, *arguable*, or *absent* (including dimensions that are simply not a primary evaluation target for that benchmark). Read across, prior work holds important *subsets* — sealed proper-scored forecasting (ForecastBench), a persistent individual (PersonaMem, Generative Agents, KnowMe-Bench), latent-goal inference (EgoToM, ToMnet, goal/plan recognition), novel-task adaptation (ARC-AGI). To our knowledge none combines all five in one prospective apparatus. The TargetSpace row reflects its proposed

design, not completed empirical validation; the table is a structured literature comparison, not a measured result.

Table 11: The five-dimension conjunction across neighboring benchmarks (1: persistent individual / evolving entity; 2: longitudinal, continuously updated evidence; 3: strictly prospective / sealed prediction; 4: calibrated proper scoring with meaningful baselines; 5: goal-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	**Persistent individual**	**Longitudinal evidence**	**Prospective / sealed**	**Calibrated + baselines**	**Transition target**
EgoToM [27]	absent	partial	partial	absent	arguable
ToMnet [6]	partial	partial	partial	absent	arguable
ForecastBench [28]	absent	absent	clear	clear	absent
PersonaMem [7]	clear	partial	absent	absent	arguable
Generative Agents, 1,000 People [5]	clear	partial	absent	absent	absent
KnowMe-Bench [26]	clear	partial	absent	absent	absent
Goal / plan recognition [29]	partial	partial	arguable	absent	arguable
ARC-AGI [65,66]	absent	absent	clear	absent	absent
TargetSpace (this work)	**clear**	**clear**	**clear**	**clear**	**clear**

12 Discussion

What a positive result establishes. The capability at stake is understanding the specific target; target-specific forecasting is its measurable proxy, and a model that cannot anticipate a target beyond its crowd and its routine does not, operationally, understand it. A positive TargetSpace result establishes calibrated, prospective, target-specific predictive skill — not understanding in a rich sense, and not the superiority of any architecture. Establishing that skill begins the scientific question rather than ending it: once a system beats R1/R2 and fails permutation, a second task opens — *explanatory audit* of which evidence, constraints, attention shifts, routine breaks, evidence tier, or belief-state updates made the forecast possible (Section 5.1). In this sense TargetSpace is forecast-first, not forecast-only: the prospective gate disciplines the explanation, so a post-hoc narrative earns credit only by improving later sealed forecasts or surviving pre-registered ablation, not by persuasion.

What it does not establish. Skill measured under passive observation is predictive, not causal: it does not identify the effect of any intervention, which would need a separate interventional design. The benchmark keeps three capabilities distinct rather than collapsing them — attention-conditioned target formation (Section 1.3), action-conditioned consequence prediction (Section 5.4), and planning — and scores only whether the predicted transition resolves correctly. A forecast can also optimize a state while erasing the context that gives it meaning (e.g. engagement at the expense of well-being);

the permutation gate and the observe-not-intervene rule are partial guards, but detecting such context erasure in general remains open.

Limitations, ethics, and deployment boundaries. (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 [13] and heterogeneous, so reporting is per-instance. (L5) Capture can be reactive — being observed may alter the behaviour being forecast — 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 forecaster is dual-use, and the observe-not-intervene rule binds the benchmark, not downstream deployers. We specify informed revocable consent, federation so raw data never leaves the target's control, aggregate-only reporting, institutional review before any human-subjects deployment, respect for recording law, and a prohibition on evaluating systems that act on the target during the window — a benchmark that rewarded changing the target would measure influence, not understanding; bystander and third-party consent, non-anonymizable multimodal streams, withdrawal of already-sealed forecasts, manipulation, and surveillance normalization remain unresolved. We keep **benchmark validity strictly separate from deployment legitimacy**: a high score certifies only better prospective predictions about a consenting individual under sealed conditions — it grants no privileged access to a person's inner life, is distinct from diagnosis, and confers no licence for employment, insurance, surveillance, coercive, or manipulative use. We model the lived trajectory only through its observable longitudinal traces and score only forecasts of future observable states; the framework claims nothing about subjective experience, qualia, or consciousness, and a high score is evidence of predictive alignment with a target, not of access to a person's mind. A system with rich observation and persistent memory must stay objective-bounded; a containment/refusal axis alongside the skill axis is required future work. A default minimal safe pilot configuration — consenting adults, local-first storage, export of sealed forecasts and aggregate metrics only, capture exclusions, and ethics review before recruitment — is given in Appendix C (Table 13, Figure 3).

Latent world models and transfer. Latent predictors are a natural architecture class here, and their open evaluation challenges are why TargetSpace scores calibrated distributions against sealed external outcomes rather than internal compatibility: latent state can be hard to ground, deterministic prediction may under-represent uncertainty, and long-horizon rollouts can drift [76]. A coherent internal rollout is not itself evidence of competence. Whether target-specific structure is reusable across regime shifts, beyond routine and memorization, is a transfer question the same bits-Skill and splits can measure (Section 13).

13 Open Research Questions

Each is a controlled comparison of target-specific Skill, grouped into five programs. **1. Evidence sufficiency and modality value:** which evidence tier first beats R2, per task and domain, and which modalities add marginal skill? **2. Attention and target-state formation:** does attention-trajectory evidence improve calibrated forecasts beyond routine and stated goals (attention-conditioned lift), and when does new evidence *create* a previously absent target rather than reveal

one — how early can an emerging target be detected before explicit commitment (time-to-correct-resolution)? **3. Transfer beyond routine and memorization:** does target-specific structure stay predictive across routine-breaking, context-shifted, and withheld windows, distinct from memorization, and how fast does skill decay once evidence stops? **4. Architecture, consequence prediction, and horizon:** do systems with an explicit consequence model calibrate better than reactive policies; at what horizon do detailed rollouts become less useful than abstract target-state forecasts; and can TargetSpace distinguish genuine consequence modelling from policy imitation? **5. Reactivity, ethics, and active evidence acquisition:** can target-state formation be separated from observer-induced reactivity, and can a system recognize when it lacks enough evidence and request more without overreaching — turning evidence acquisition into a measured, consent-aware, privacy-bounded capability? Recent work frames reusable, computationally-bounded structure as *epiplexity* [80]; we note it as related framing but do not adopt it as a metric, since estimating it needs a reference learner's training dynamics that closed APIs, frozen models, and human baselines do not expose, whereas the bits-Skill above already reports the prospective predictive code-length gain in an architecture-neutral way.

14 Conclusion

Many systems now claim to understand specific targets, and almost none can show their forecasts are target-specific rather than reflections of base rates and routine. TargetSpace makes that distinction measurable: prospective, sealed, proper-scored forecasting of target-state transitions under partial observation, certified by an own-routine baseline, a calibration gate, and a permutation specificity test, with an evidence-tier ablation and an architecture-neutral grid. Its novelty is the conjunction assembled around one question — *is the forecast about the target, or about the average?* — not any single ingredient, each of which it adopts and cites. Personal world modeling is the flagship instantiation; some target states are not fixed labels but form through attention and action, so longitudinal attention is both evidence for the forecast and part of the process forecast, and a capable system must forecast consequences across horizons — calibrated and target-specific — whether reactive, predictive, or hybrid. TargetSpace scores that external forecast, not the architectural story. It begins from one premise: the target was never a profile, a preference list, or a memory store, but a lived trajectory — an unfolding stream of attention, pressure, constraint, and emerging goals — which the benchmark asks a model to infer from observable traces and to forecast where it turns next.

Forecast the target, not the average — and model how the target is being formed, not only which label best describes it.

The target is not a profile. The target is a lived trajectory.

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 availability. A minimal synthetic demonstration harness accompanies this preprint as an ancillary file (Appendix A); it uses no participant data and supports no empirical claims. The full reference implementation and participant-data harness remain in preparation.
Project website. The project website is in preparation and is not cited in this version.
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.
[16] C. White et al. LiveBench: A Contamination-Limited LLM Benchmark. 2024. arXiv:2406.19314.
[17] SWE-bench-Live: Contamination-Resistant Evaluation for Software Agents. 2025. arXiv:2505.23419.
[20] A. Wang et al. GLUE: A Multi-Task Benchmark for Natural Language Understanding. ICLR, 2019. arXiv:1804.07461.
[21] M. Chen et al. Evaluating Large Language Models Trained on Code (HumanEval). 2021. arXiv:2107.03374.
[43] P. Liang et al. Holistic Evaluation of Language Models (HELM). TMLR, 2023. arXiv:2211.09110. (Calibration as a first-class measured axis.)
[65] F. Chollet. On the Measure of Intelligence. 2019. arXiv:1911.01547.
[66] 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

[13] A. Bellot, J. Richens, T. Everitt. The Limits of Predicting Agents from Behaviour. ICML, 2025. arXiv:2506.02923.
[91] C. Song, Z. Qu, N. Blumm, A.-L. Barabasi. Limits of Predictability in Human Mobility. Science, 327(5968):1018–1021, 2010.
[14] I. J. Good. Rational Decisions. J. Royal Statistical Society B, 14(1), 1952.
[15] T. Gneiting, A. E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. JASA, 102(477), 2007.
[18] G. W. Brier. Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review, 78(1), 1950.
[22] A. H. Murphy. What Is a Good Forecast? Weather and Forecasting, 8(2), 1993.
[28] E. Karger et al. ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities. ICLR, 2025. arXiv:2409.19839.
[29] S. Keren, A. Gal, E. Karpas. Goal Recognition Design (worst-case distinctiveness). ICAPS, 2014.
[40] Q. Yang et al. Prophet Arena (LLM-as-a-Prophet): a live prediction-market forecasting benchmark. 2025. arXiv:2510.17638.

[80] M. Finzi, S. Qiu, Y. Jiang, P. Izmailov, J. Z. Kolter, A. G. Wilson. From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence. 2026. arXiv:2601.03220.

C. World models, JEPA, predictive learning

[12] D. Ha, J. Schmidhuber. World Models. NeurIPS, 2018. arXiv:1803.10122.
[19] K. Friston. The Free-Energy Principle: A Unified Brain Theory? Nat. Rev. Neuroscience, 11, 2010.
[32] C. Heins et al. pymdp: A Python library for active inference. JOSS, 2022. arXiv:2201.03904.
[33] Y. LeCun. A Path Towards Autonomous Machine Intelligence. v0.9.2, 2022. OpenReview BZ5a1r-kVsf. (Position paper; not arXiv.)
[34] M. Assran et al. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture (I-JEPA). CVPR, 2023. arXiv:2301.08243.
[35] A. Bardes et al. Revisiting Feature Prediction for Learning Visual Representations from Video (V-JEPA). 2024. arXiv:2404.08471.
[36] M. Assran et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. 2025. arXiv:2506.09985.
[37] D. Hafner et al. Mastering Diverse Domains through World Models (DreamerV3). 2023. arXiv:2301.04104.
[38] J. Schrittwieser et al. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model (MuZero). Nature 588, 2020. arXiv:1911.08265.
[39] R. P. N. Rao, D. H. Ballard. Predictive Coding in the Visual Cortex. Nature Neuroscience, 2(1), 1999.
[41] A. Warrier et al. Benchmarking World-Model Learning with Environment-Level Queries (AutumnBench / WorldTest). 2025. arXiv:2510.19788.
[42] D. Chen et al. WorldPrediction: A Benchmark for High-level World Modeling and Long-horizon Procedural Planning. 2025. arXiv:2506.04363.
[44] A. van den Oord, Y. Li, O. Vinyals. Representation Learning with Contrastive Predictive Coding (CPC). 2018. arXiv:1807.03748.
[45] A. Bardes, J. Ponce, Y. LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. ICLR, 2022. arXiv:2105.04906.
[46] G. Zhou, H. Pan, Y. LeCun, L. Pinto. DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning. 2024. arXiv:2411.04983.
[47] J. Schmidhuber. Making the World Differentiable: On Using Fully Recurrent Self-Supervised Neural Networks for Dynamic Reinforcement Learning and Planning. Tech. Rep. FKI-126-90, TU Munich, 1990.
[48] H. B. Barlow. Possible Principles Underlying the Transformation of Sensory Messages. In Sensory Communication (W. Rosenblith, ed.), MIT Press, 1961.
[49] A. Radford et al. Learning Transferable Visual Models From Natural Language Supervision (CLIP). ICML, 2021. arXiv:2103.00020.
[50] A. Brohan et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. 2023. arXiv:2307.15818.
[51] K. Black et al. (Physical Intelligence). $\pi 0$: A Vision-Language-Action Flow Model for General Robot Control. 2024. arXiv:2410.24164.
[76] N. O. Lambert, K. S. J. Pister, R. Calandra. Investigating Compounding Prediction Errors in Learned Dynamics Models. 2022. arXiv:2203.09637.
[77] D. Hafner, K.-H. Lee, I. Fischer, P. Abbeel. Deep Hierarchical Planning from Pixels. NeurIPS, 2022. arXiv:2206.04114.

[78] Physical Intelligence. $\pi_{0.7}$: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities. 2026. arXiv:2604.15483.
[79] D. Chen et al. VL-JEPA: Joint Embedding Predictive Architecture for Vision-Language. 2025. arXiv:2512.10942.

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

[3] A. Salemi et al. LaMP: When Large Language Models Meet Personalization. SIGIR, 2024. arXiv:2304.11406.
[4] J. S. Park et al. Generative Agents: Interactive Simulacra of Human Behavior. UIST, 2023. arXiv:2304.03442.
[5] J. S. Park et al. Generative Agent Simulations of 1,000 People. 2024. arXiv:2411.10109.
[6] N. C. Rabinowitz et al. Machine Theory of Mind (ToMnet). ICML, 2018. arXiv:1802.07740.
[7] B. Jiang et al. Know Me, Respond to Me: Benchmarking LLMs for Dynamic User Profiling (PersonaMem). COLM, 2025. arXiv:2504.14225.
[8] S. Zhao et al. Do LLMs Recognize Your Preferences? (PrefEval). ICLR, 2025. arXiv:2502.09597.
[9] Twin-2K-500: Digital Twins of over 2,000 People. 2025. arXiv:2505.17479.
[10] J. Li et al. How Far are LLMs from Being Our Digital Twins? (BehaviorChain). Findings of ACL, 2025. arXiv:2502.14642.
[11] M. Binz, E. Schulz et al. A Foundation Model to Predict and Capture Human Cognition (Centaur). Nature, 2025. arXiv:2410.20268.
[26] T. Wu et al. KnowMe-Bench: Benchmarking Person Understanding for Lifelong Digital Companions. 2026. arXiv:2601.04745.
[27] EgoToM: Benchmarking Theory of Mind Reasoning from Egocentric Videos. Meta, 2025. arXiv:2503.22152.
[31] MMToM-QA: Multimodal Theory of Mind Question Answering. ACL, 2024. arXiv:2401.08743.
[56] K. Grauman et al. Ego4D: Around the World in 3,000 Hours of Egocentric Video. CVPR, 2022. arXiv:2110.07058.
[57] K. Grauman et al. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. CVPR, 2024. arXiv:2311.18259.
[23] R. E. Nisbett, T. D. Wilson. Telling More Than We Can Know: Verbal Reports on Mental Processes. Psychological Review, 84(3):231–259, 1977.
[24] H. A. Simon. Designing Organizations for an Information-Rich World, in M. Greenberger (ed.), 1971. (Attention as the scarce resource.)
[58] 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.
[59] B. S. Connelly, D. S. Ones. An Other Perspective on Personality: Meta-Analytic Integration of Observers' Accuracy and Predictive Validity. Psychological Bulletin, 136(6):1092–1122, 2010.
[60] S. Vazire, M. R. Mehl. Knowing Me, Knowing You: The Accuracy and Unique Predictive Validity of Self-Ratings and Other-Ratings of Daily Behavior. J. Personality and Social Psychology, 95(5):1202–1216, 2008.
[61] S. Shiffman, A. A. Stone, M. R. Hufford. Ecological Momentary Assessment. Annual Review of Clinical Psychology, 4:1–32, 2008.
[62] 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.
[63] J. Carden, R. J. Jones, J. Passmore. Defining Self-Awareness in the Context of Adult Development: A Systematic Literature Review. J. Management Education, 46(1):140–177, 2022.

[64] J.-P. Onnela, S. L. Rauch. Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. Neuropsychopharmacology, 41(7):1691–1696, 2016.
[92] T. Nagel. What Is It Like to Be a Bat? The Philosophical Review, 83(4):435–450, 1974.

E. Cross-scale target-state systems

[25] M. Levin. Technological Approach to Mind Everywhere (TAME). Frontiers in Systems Neuroscience, 16:768201, 2022.
[30] M. Lange et al. CellRank for directed single-cell fate mapping. Nature Methods, 19, 2022.
[67] D. D. Garrett, G. R. Samanez-Larkin, S. W. S. MacDonald, U. Lindenberger, A. R. McIntosh, C. L. Grady. Moment-to-Moment Brain Signal Variability: A Next Frontier in Human Brain Mapping? Neuroscience & Biobehavioral Reviews, 37(4):610–624, 2013.
[68] A. A. Faisal, L. P. J. Selen, D. M. Wolpert. Noise in the Nervous System. Nature Reviews Neuroscience, 9(4):292–303, 2008.
[69] G. M. Edelman, J. A. Gally. Degeneracy and Complexity in Biological Systems. Proceedings of the National Academy of Sciences, 98(24):13763–13768, 2001.
[70] A. A. Prinz, D. Bucher, E. Marder. Similar Network Activity from Disparate Circuit Parameters. Nature Neuroscience, 7(12):1345–1352, 2004.

F. Domain forecasting precedents (track validation regimes)

[52] T. Hong et al. Probabilistic Energy Forecasting: Global Energy Forecasting Competition 2014 (GEFCom2014). Int. J. Forecasting, 32(3), 2016.
[53] M. A. Reyna et al. Early Prediction of Sepsis (PhysioNet/Computing in Cardiology Challenge 2019). Critical Care Medicine, 2020.
[54] C. Marling, R. Bunescu. The OhioT1DM Dataset and the Blood Glucose Level Prediction (BGLP) Challenge. KDH/KHD, 2018/2020.
[55] X. Xu et al. GLOBEM: Generalization of Longitudinal Behavior Modeling. NeurIPS Datasets & Benchmarks, 2022.

G. Attention and goal dynamics

[71] E. I. Knudsen. Fundamental Components of Attention. Annual Review of Neuroscience, 30:57–78, 2007.
[72] M. Corbetta, G. L. Shulman. Control of Goal-Directed and Stimulus-Driven Attention in the Brain. Nature Reviews Neuroscience, 3(3):201–215, 2002.
[73] A. Dijksterhuis, H. Aarts. Goals, Attention, and (Un)Consciousness. Annual Review of Psychology, 61:467–490, 2010.
[74] W. Ocasio. Towards an Attention-Based View of the Firm. Strategic Management Journal, 18(S1):187–206, 1997.
[75] K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, J. O'Doherty, G. Pezzulo. Active Inference and Learning. Neuroscience & Biobehavioral Reviews, 68:862–879, 2016.

H. Memory, belief states, and goal inference

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

[82] D. L. Schacter, D. R. Addis. The Cognitive Neuroscience of Constructive Memory: Remembering the Past and Imagining the Future. Philosophical Transactions of the Royal Society B, 362(1481):773–786, 2007.
[83] E. Tulving. Episodic and Semantic Memory. In E. Tulving, W. Donaldson (eds.), Organization of Memory, pp. 381–403. Academic Press, New York, 1972.
[84] P. Sheeran, T. L. Webb. The Intention–Behavior Gap. Social and Personality Psychology Compass, 10(9):503–518, 2016.
[85] D. C. McClelland, R. Koestner, J. Weinberger. How Do Self-Attributed and Implicit Motives Differ? Psychological Review, 96(4):690–702, 1989.
[86] C. L. Baker, R. Saxe, J. B. Tenenbaum. Action Understanding as Inverse Planning. Cognition, 113(3):329–349, 2009.
[87] C. L. Baker, J. Jara-Ettinger, R. Saxe, J. B. Tenenbaum. Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human Mentalizing. Nature Human Behaviour, 1:0064, 2017.
[88] 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.
[89] B. Alderson-Day, C. Fernyhough. Inner Speech: Development, Cognitive Functions, Phenomenology, and Neurobiology. Psychological Bulletin, 141(5):931–965, 2015.
[90] P. Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS, 2020. arXiv:2005.11401.

A Minimal synthetic harness

Purpose. The harness is a sanity check of the evaluation machinery, not empirical evidence that passive observation improves personal forecasting. It exercises the input and output schemas, the two baselines, the scoring rules, and the diagnostic gates end to end on data we generate ourselves, so that a reader can confirm the pipeline runs and produces the reported quantities. The shipped artifact is a single deterministic, pure-Python-stdlib script distributed as an arXiv ancillary file (`anc/targetspace_synthetic_demo.py`). The multi-file layout below is the planned structure of the fuller reference implementation, in preparation.

What it demonstrates. Synthetic target histories; walk-forward forecast instances; the R1 population-prior baseline; the R2 own-routine baseline; a simulated model forecast; log (in bits) and Brier scoring; a calibration diagnostic; a permutation specificity check; and evidence-tier reporting.

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.

Planned file layout (reference implementation, in preparation).

- `README.md`
- `generate_synthetic_targets.py`
- `run_walk_forward_eval.py`
- `baselines.py`
- `scoring.py`

- permutation_test.py
- example_config.yaml
- synthetic_results_example.csv
- leaderboard_example.csv

Forecast-instance (input) schema. instance_id, target_id, forecast_time, resolution_time, question_type, answer_space, evidence_tier, available_features, resolved_outcome, split_or_fold, admissible_for_R2.

Forecast-output schema. forecast_id, system_id, target_id, forecast_time, question_type, answer_space, probability_vector, evidence_tier, model_access_mode, sealed_hash, timestamp, adapter_version.

Example commands. python generate_synthetic_targets.py -config example_config.yaml ; python run_walk_forward_eval.py -config example_config.yaml ; python permutation_test.py -predictions outputs/predictions.csv.

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

Table 12: 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

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 forecast counts.

B 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-forecast routine features: marginal historical frequency for the target × 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 forecast 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 forecast 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.

C Safe pilot configuration

The benchmark can be tested without assembling an uncontrolled lifelogging dataset. Table 13 lists a default minimal safe pilot configuration, and Figure 3 shows the federated, local-first architecture it assumes: raw capture stays on the participant device, and only sealed forecasts, resolved labels, and aggregate metrics leave the client.

D Supplementary figures

These figures restate, in visual form, material presented as tables in the main text: Figure 4 corresponds to the target-specificity stack of Section 3, and Figure 5 to the evidence ladder of Section 6.2.

Table 13: 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 forecasts 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 redaction or exclusion where feasible.
• 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.

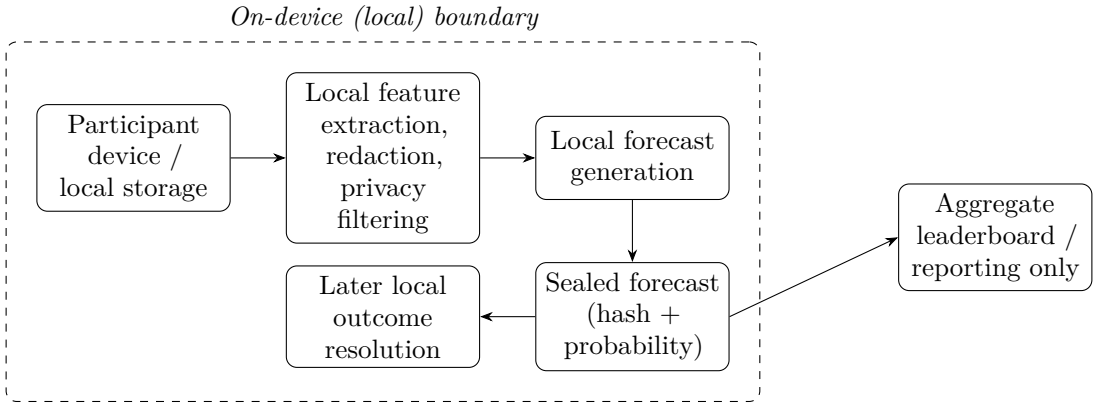

Figure 3: Safe federated pilot architecture. Raw audio, video, and screen data stay on the participant device: feature extraction, redaction, privacy filtering, forecast generation, sealing (hash and probability), and outcome resolution all run locally within the dashed on-device boundary. Only the sealed forecast (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.

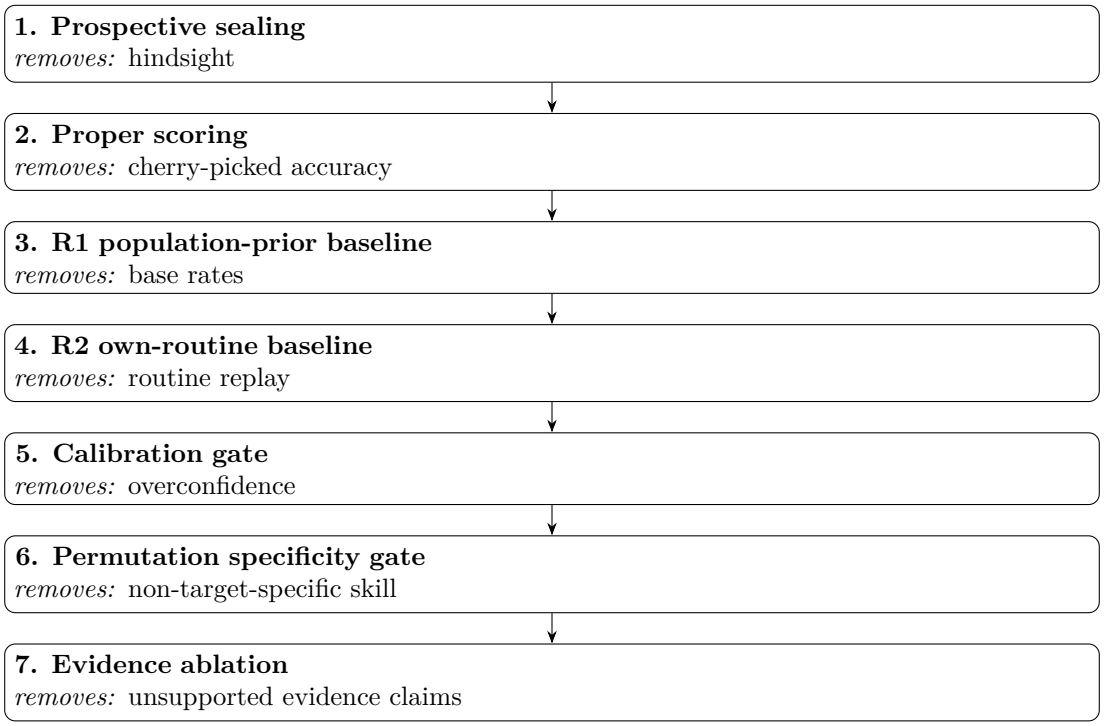

Figure 4: The target-specificity stack, complementing the stack table (Section 3). Each layer removes one way a forecast can appear target-specific without being so: a forecast that survives all seven filters cannot be explained by hindsight, selective reporting, population base rates, replay of the subject's own routine, overconfidence, non-target-specific skill, or unsupported evidence claims. Layers 4 and 6 (the R2 own-routine baseline and the permutation specificity gate) are the anchors contributed here.

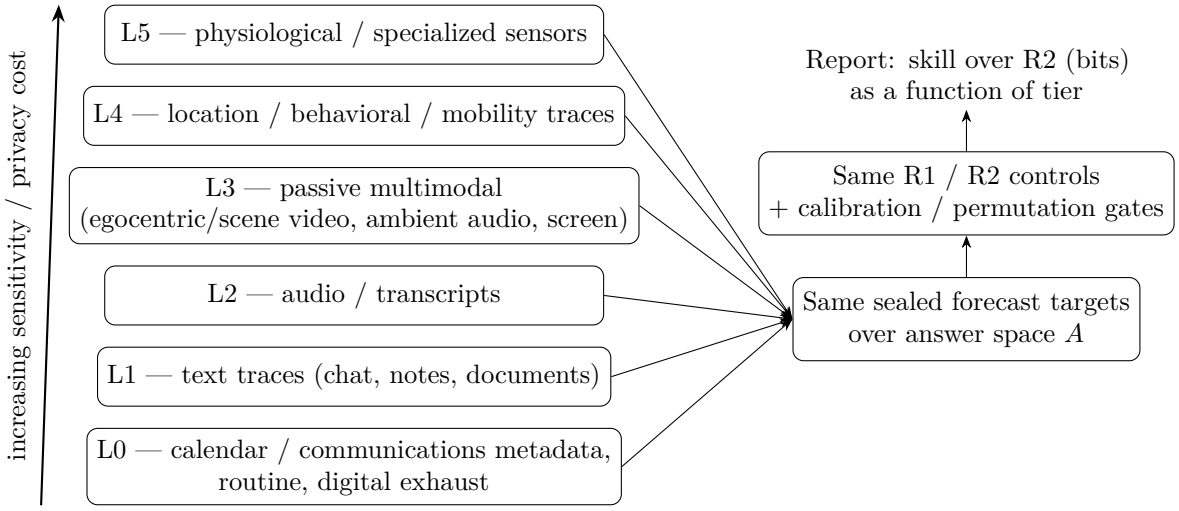

Figure 5: Evidence-tier ablation design (Section 6.2). Six evidence tiers (L0, least sensitive, at bottom; L5, most sensitive, at top) are each evaluated against the *same* sealed forecast targets and the *same* R1 (population-prior) and R2 (own-routine) controls under the calibration and permutation-specificity gates. Whether richer evidence helps is measured as skill over R2 (in bits) as a function of tier, not assumed; sensitivity and privacy cost increase upward.