DECODE: DECOMPOSABLE ENVIRONMENTS FOR CONTROLLED SEQUENTIAL-DECISION EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Benchmarks for sequential decision systems usually report a final score, which shows that a solver fell short of the optimum but not why. DECODE is a generator of partially observed sequential decision environments whose equations are known. This allows the gap between a solver and the exact optimum to be split into components, each measured against a reference computed from those equations in closed form, analytically, or by Monte Carlo with certified precision. The components are the planning horizon, knowledge of the true action response, exact inference over the hidden state, memory, re-observation before each decision, and the inflation of the solver's own value estimate. The generator is instantiated as a consumer-credit collections portfolio and released as five environments that share the same randomly drawn dynamics matrices and differ in a memory parameter, which sets how much an action's effect depends on the client's history, and a noise parameter. In the deterministic environment, with both parameters at zero, where scores are percentages of the exact optimum, planning six months ahead instead of one raises the same solver's score by 15 to 17 percentage points, while switching among the tested solver classes at a fixed horizon changes it by at most 4. The best solver trained on the logs ends 15 points below the optimum. Giving it the true action response together with exact inference recovers 9 of these points, and within that disclosure exact inference is worth nothing measurable. The missing information is linearly recoverable from the logged observables, so the binding constraint is the lack of exploration in the logged data, which none of the tested estimator classes removes. On the same environment, a planner acting on expected future states overstates the value of its own plan even with exact beliefs, so self-reported values cannot replace external evaluation. Re-observing before each decision is worth almost nothing in the deterministic environment and gains value as process noise grows. When the memory parameter is low, all tested recurrent configurations perform the same, so such environments cannot rank memory architectures. A statistic that needs no ground truth relates the released environments to public Fannie Mae loan-performance data. The generator, the five environments, and all artifacts behind the reported values are released with the paper.

1 INTRODUCTION

Usually, when benchmarks rank the solvers of a data-driven sequential decision problem, the explanation of a solver's gap to the optimum is not supplied and requires quantities the environment does not provide. A practitioner reading such a score learns who won but cannot tell whether the gap comes from a horizon chosen too short, an estimator failing to extract what the logs contain, or logs that never contained the information.

DECODE is a generator of partially observed sequential decision environments built so that those quantities exist. Each environment is a finite-horizon partially observable Markov decision process (POMDP) (Kaelbling et al., 1998) with time-invariant dynamics, instantiated as a consumer-credit collections portfolio in which a hidden account state evolves under known dynamics, monthly actions perturb it through known response maps (the action-dependent part of the transition and repayment model), and value is realized as recovered debt net of contact costs (the reward). A solver's belief is the probability distribution of the hidden state given everything it has observed. DECODE

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

serves as a measurement instrument, the generator together with the references it computes. Because the world's equations are known, each component of a solver's gap that the instrument isolates carries a reference, and the references form a ladder that survives as the physics hardens, from closed form through path-dependent analytic values to Monte Carlo with certified precision, with a no-fit rung that separates the information a log contains from what any estimator extracts. Solvers read only what a deployed system would log, while labeled reference solvers never compete and receive controlled disclosures of ground truth, whose value is read from the graded solver's response. Two physics parameters, a memory parameter and process noise, set how much memory and re-observation are worth, and five settings of the two are released as "members", each passing machine-checked acceptance checks before release (Fig. 1).

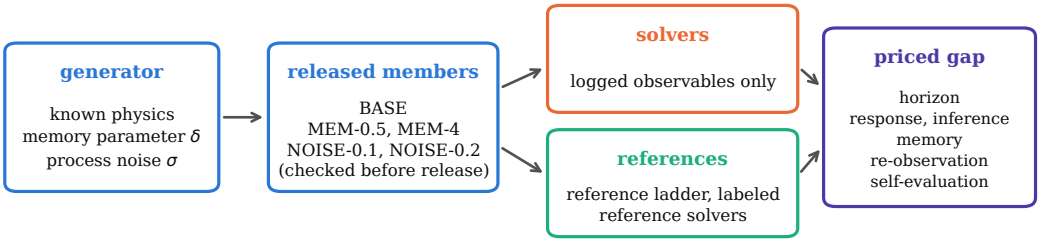

Figure 1: Overview of DECODE.

On the member with both parameters at zero (BASE), planning six months ahead instead of one is worth 15 to 17 points of the exact optimum for the same solver, whereas switching between the tested solver classes at a fixed horizon never moves the score by more than 4 points. The best solver fitted on the logs alone ends 15 points below the optimum. Handing it the world's true response to each action together with exact inference over the hidden state recovers 9 of them, and within that disclosure exact inference changes nothing measurable. A diagnostic that reads the hidden state directly then finds the missing information linearly recoverable from the logged observables, so what limits the solver is how the data were collected, with almost no variation in the actions, a limit none of the tested estimator classes removes.

On worlds where a client's history carries little information, every tested recurrent configuration performs the same, so results obtained there cannot rank memory architectures. And a planner that values its own plan by chaining expected states overstates that value even when its beliefs are exact, so a planner's self-report cannot replace external grading.

This paper contributes the instrument DECODE, a priced decomposition of solver gaps on BASE, measurements of memory and re-observation value across the released members, and a transport experiment relating those worlds to a public loan-performance portfolio through a statistic computable without ground truth. The generator, the five members, and every artifact behind the reported values are in the anonymized reproduction package of the supplementary material, described in the reproducibility statement.

Suites that carry a computed optimum report performance against it, including along difficulty ladders (Osband et al., 2020; Rajan et al., 2023; Ibrahim et al., 2026), and the closest decomposition prices estimator-internal errors on fixed domains (Fu et al., 2019). DECODE combines a reference recomputed at every rung of a hardening ladder with a same-task gap decomposed into informational ingredients (Section 5).

2 DECODE

A world is a population of clients whose monthly outcomes depend on a hidden state that evolves from one month to the next. In the example presented in this paper, consumer debt collection, the clients are delinquent loans, the outcomes are repayments, and the actions are contact choices and settlement offers, each at a cost. A solver never accesses the hidden state. It observes the public account state and a noisy reading of the hidden state, and chooses an action per client per month. Two parameters set the physics, the memory parameter δ and the process-noise scale σ. Five members are released, BASE with both parameters at zero, MEM-0.5 and MEM-4 with $\delta = 0.5$

and 4 and no process noise, and NOISE-0.1 and NOISE-0.2 with $\sigma = 0.1$ and 0.2 and $\delta = 0$ (Table 3 in Appendix B). Each term is defined where it first appears and is collected in the glossary of Appendix A.

2.1 PHYSICS

Each client carries a hidden state $z_t \in \mathbb{R}^6$ drawn from a standard Gaussian population at origination and evolving once per month. An action $a_t = (c_t, d_t)$ pairs a contact channel c_t, one of none, SMS, WhatsApp, or phone, with a settlement offer d_t. The channels carry contact intensities $w(c) = 0$, 0.4, 0.4, and 1 respectively, so that SMS and WhatsApp move the state equally, at equal cost, and differ only in their effect on repayment. The state transition, a linear-Gaussian state-space model (Kalman, 1960) with an action-dependent rotation, is given by:

$$z_{t+1} \;=\; A\,Q_{w(c_t)}\,z_t \;+\; b\,w(c_t) \;+\; \eta_t, \qquad \eta_t \sim \mathcal{N}(0, \sigma^2 I), \tag{1}$$

where A is a contraction, b a fixed push direction, and $Q_w = \exp(w\,\delta\,S)$ a rotation generated by a fixed skew-symmetric S. The memory parameter δ scales the rotation. At $\delta = 0$, Q_w is the identity and every action adds the same push $b\,w(c)$, so the contributions of past actions add up independently and their order matters only through how recently each was taken. At $\delta > 0$ each action also rotates the current state, and since rotating and contracting do not commute, the monthly maps $A\,Q_w$ for different intensities do not commute in general. Each action's effect then depends on the state and on every earlier action, and this interaction is what δ controls.

The second parameter, the process-noise scale σ, injects monthly innovations. When $\sigma = 0$ the state is a deterministic function of the client's initial draw and the action history. At $\sigma > 0$ the state becomes random, but each innovation is shrunk by A in every later month, so the accumulated variance converges to a finite level, at most $\sigma^2/(1 - \|A\|_2^2)$ per component of the hidden state.

Each month, the solver receives a two-dimensional linear reading of the hidden state, the measurement equation of that state-space model, written as:

$$o_t \;=\; C\,z_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,\, 0.6^2 I_2), \tag{2}$$

where $C \in \mathbb{R}^{2\times 6}$ is a fixed loading. Besides o_t, the solver observes the public account state x_t (delinquency ageing, balance, contact fatigue, demographics), and the logged records also carry realized repayments. Repayment is a monthly Bernoulli draw, $r_t \sim \mathrm{Bernoulli}(\mathrm{expit}(y_t))$ with expit the logistic sigmoid, whose logit, as in logistic regression (Cox, 1958), is additive in the grouped terms given by:

$$y_t \;=\; \beta_0 + \gamma_{c_t} + g(x_t) + h_{c_t}(d_t) - \phi_{c_t}(x_t)\,f_t + \kappa\,u_{c_t}^{\top} z_t + \theta_{c_t}^{\top}\tau_i + \xi_t, \tag{3}$$

where β_0 is the logit of the member's base rate, γ_c a channel offset, g collects the public-state effects, h_c is a saturating response to the discount d_t, and f_t is the contact-fatigue component of x_t, with a sensitivity $\phi_c(x_t)$ that depends on the borrower's observed income. Both h_c and ϕ_c are the same for every contact channel and vanish on no-contact months. The term $\kappa\,u_{c_t}^{\top} z_t$ is the hidden score of the acting channel c_t, with coupling vectors u_c and scale κ. The term τ_i holds four hidden traits of borrower i (resilience, payment friction, regional economic stress, and contactability) with fixed weights θ_c. Finally, $\xi_t \sim \mathcal{N}(0, 0.2^2)$ is noise drawn per loan-month. Two of the hidden traits, payment friction and contactability, also steer the logging policy, the rule that generated the historical records, toward phone contact and away from no contact, so they confound treatment and outcome. Neither the traits nor ξ_t can be recovered from any information set short of the oracle, the grader's side of the world, which holds the hidden state and the equations.

A "geometry" is one random draw, shared by every client of a world, of the fixed matrices and vectors in its equations, namely A and b in the transition (Eq. 1), the loading C of the reading (Eq. 2), the coupling vectors u_c of the hidden score (Eq. 3), and, on members with $\delta > 0$, the rotation generator S. The released members share one geometry, and robustness sweeps use four further draws.

2.2 PROTOCOL

Worlds are released in train–evaluation pairs drawn with independent seeds under one physics. A training world of one million borrowers over 24 months is logged under the logging policy, and an

evaluation world of one hundred thousand borrowers over 12 months supplies evaluation cohorts. On eligible loan-months the logging policy substitutes, with probability $\varepsilon = 0.03$, one of the three contact channels, drawn uniformly, together with a uniformly drawn discount. The no-contact action is never explored. Months zero through five are generated under the logging policy and handed to every solver as the loan's public history. The decision point is month six, and the solver then decides every month over the remaining six-month horizon. Every solver is graded on the same fixed cohort of 4,000 loans at month six. Under process noise the cohort is 2,000 loans, each followed along 24 innovation paths shared by all solvers, sized to the precision bar of Section 2.3. Grading is closed-loop, meaning that each solver re-decides every month under counterfactual observations of its own trajectory, generated with common random numbers so that paired comparisons cancel shared noise.

A month's value is the cash the solver's action is expected to recover, given the repayment probability of Eq. 3. That expected cash is weighted by the probability that the loan is still open, and the contact cost is subtracted. No payment is drawn during grading, so the solver never learns whether a client paid and keeps receiving the readings and the account's deterministic updates, such as delinquency ageing and contact fatigue. Each action pairs one of the four channels with one of 21 discount levels from zero to half the balance, $(4 \times 21)^6$ six-month plans in all.

2.3 Reference ladder

A "disclosure" replaces one input of a graded solver by its ground truth, such as the true response to each action or the exact belief about the hidden state. The copy that receives it is a "reference solver", labeled as such, and it never competes with the graded solvers. The value of a disclosure is the resulting change in the graded solver's value. It is a property of that solver and can be negative when the solver's remaining inputs are biased. It therefore differs from the decision-theoretic value of information (Howard, 1966), which an optimal maximizer can never find negative.

At $\sigma = 0$ the "exact optimum" is computed with the true hidden state at the decision point. The plan then fixes the state path, and since the discount leaves the hidden state unchanged, every plan is priced exactly by enumerating the 4^6 channel sequences and solving the discount grid within each by backward dynamic programming. A one-month optimum computed the same way isolates the value of the horizon.

The "exact filter" computes what the readings reveal about the hidden state. Given every reading and action of a client up to month t, the belief about z_t is Gaussian, since the model is linear with Gaussian noise, and its mean and covariance follow the Kalman recursion (Kalman, 1960) in closed form on every member, including under process noise. The belief is exact for that information and ignores repayment outcomes. A planner on this exact belief bounds the filtering value of memory among solvers that chain certainty-equivalent cell probabilities, treating current estimates as truth, with future readings masked. Such a planner is passively adaptive, neglecting any dual effect of actions on future information (Bar-Shalom & Tse, 1974). A solver that integrates over the joint belief of the hidden states along a path is admissible and may exceed this reference.

Memory is the use of a client's past readings and actions to infer its hidden state. Three probe scores measure how much outcome-relevant information an information set carries (Table 1), and memory value is what a set with history scores above one without it. Each score is the out-of-fold R^2, over three folds, of a gradient-boosted regressor (Friedman, 2001) fit to the oracle's true recovery logit y_t. The rows are the phone-contact months among the exploration rows of Section 2.2 that have a complete six-month window, so the target is one function of the hidden state and the channel is randomly assigned.

The probe scores have a ceiling that can be computed without fitting any model. The part of the logit that depends on the hidden state is the hidden score $\kappa\, u_c^\top z_t$, with c the phone channel of the measurement rows. From the readings and actions, the most that can be known about z_t is the mean μ_t of the exact belief, so even a perfect predictor leaves the error $\kappa\, u_c^\top (z_t - \mu_t)$ unexplained. Because the hidden score enters the logit linearly, the variance of that error follows from the belief, and the score of any probe that learns about z_t only from readings and actions is bounded by the "information bound", written as:

$$R^2_{\max} \;=\; 1 \;-\; \frac{\mathrm{Var}\big(\kappa\, u_c^\top (z_t - \mu_t)\big)}{\mathrm{Var}(y_t)}. \tag{4}$$

Table 1: Information probes and the two contrasts built from them.

Probe	Information set
Snapshot	Current month's public state, reading, and offered discount
History	Snapshot set plus the last six months of readings and actions (recorded by channel)
True state	Snapshot set plus the hidden state z_t
Contrast	**What it measures**
Separation, history − snapshot	Memory value a fixed six-month window extracts
Headroom, true state − history	Value of knowing the state outright over the window

The bound serves as a check. A fitted model that scores above it signals an instrument fault or information about z_t carried by the repayment history, which the belief does not use. A model that scores below it falls short through estimation error and through those unknown terms.

Under process noise, exact plan values are replaced by Monte Carlo estimates over innovation paths shared across solvers. Plans are still chosen by the noise-free dynamic program above, applied to the current belief or, for the upper reference, to the true state, and are then graded under noise. These references are values of specified policies, with no optimality guarantee in the noisy world. Each member's upper reference is the exact optimum at $\sigma = 0$ and, under noise, the value of monthly re-planning with the true state, which normalizes values and carries no certificate as a bound on attainable value. Precision is certified, since a reference is read only after every solver's Monte Carlo standard error falls below half a percent of the upper reference.

2.4 Checks

Every member passes machine-checked realism bands before release, which require the share of loans recovered within one month and within nine months to fall inside pre-set ranges, on the training and the evaluation world of the four other members and on a 100,000-borrower prefix of BASE's training world. The three members with an exact reference additionally pass pre-registered information clauses, which test that a world holds the information structure its parameters were set to produce (Appendix B). Every released member passed its checks before any result in this paper was read. Each release carries its check report and a regeneration manifest. Check criteria were fixed before generation. One information clause was amended once before release, and three members were regenerated with an adjusted base rate, all recorded with the release (Appendix B).

2.5 Solvers

The graded solvers are built from the training world's logs. All but LLM-agent fit a model of repayment, or of expected cash, as a function of the account state and the action, and turn its predictions into actions. A "myopic" solver chooses, each month, the action with the highest expected value for that month alone. A "sequential" solver plans the whole remaining horizon every month, with the same enumeration of channel sequences and the same dynamic program over discounts as the exact optimum (Section 2.3) but on its own predicted probabilities, executes the first action, and re-plans the next month.

GBM-lags, the solver to which the disclosures of Section 2.3 are applied, combines two gradient-boosted machines (GBMs) (Friedman, 2001) used as classifiers with calibrated probabilities. One predicts repayment on contact months and is fit on the exploration rows. The other predicts repayment on no-contact months and is fit on the logged no-contact months. GBM-lags reads the public state, the current reading, and the last six months of readings and actions. GBM-flat is the same engine without the six months of history. Neither engine can imagine its own future readings, so during a sequential search their reading and history features stay at their decision-month values.

CE-Seq is a sequential certainty-equivalent (CE) planner on a learned belief. A recurrent model summarizes the account's full observed history, from its first month, and the planner chains the model's predicted probabilities across the remaining horizon as if they were certain, with future readings masked. Two further solvers decide one month at a time. GBM-cash fits a gradient-boosted

model of the expected monthly cash for each state and action with a Tweedie loss, on the whole logged book. LLM-agent applies, month by month, a fixed set of state-conditional rules that a large language model (LLM) wrote after examining a censored copy of the training logs (Appendix E).

A solver class is a solver's architecture and estimator, so GBM-lags and GBM-flat, which differ only in the history they read, form one class. Only these two are graded at both horizons, so the horizon contrasts of Section 3.1 compare each of them with itself. The recurrent configurations of Section 3.2 are memory encoders that the probes score. Appendix C details each solver (Table 4).

3 RESULTS

Closed-loop values carry 95% paired bootstrap intervals over loans, with 10,000 resamples that draw the same loans for every solver. Probe scores (Table 1) are out-of-fold means over three cross-validation folds. The recurrent configurations of Section 3.2 report the mean and standard deviation over five training seeds on MEM-4 and two on the other members, and on BASE and MEM-0.5 their seeds agree within 0.01 in R^2. Monte Carlo references carry per-solver standard errors (Section 2.3). Bootstrap intervals are percentile intervals, with no correction for multiple comparisons. A point is one percentage point of the member's upper reference, and a point of R^2 is 0.01.

3.1 BASE

On BASE, within a solver class, extending the planning horizon from one month to six is worth between 15 and 17 points of the exact optimum, and no contrast between solvers at a fixed horizon exceeds 4 points. The largest, 3.75, separates GBM-lags from GBM-flat, two solvers of one class that differ only in the history they read, and the largest across fitted solver classes is 2.1 (Fig. 2, Appendix H). In Fig. 2(a), the connected pairs are these within-class horizon contrasts, unconnected markers are solvers graded at one horizon, and the dotted rule marks the one-month optimum. LLM-agent, graded once at the one-month horizon in August 2026, reaches 64.74% of the exact optimum, below every fitted solver (Appendix E).

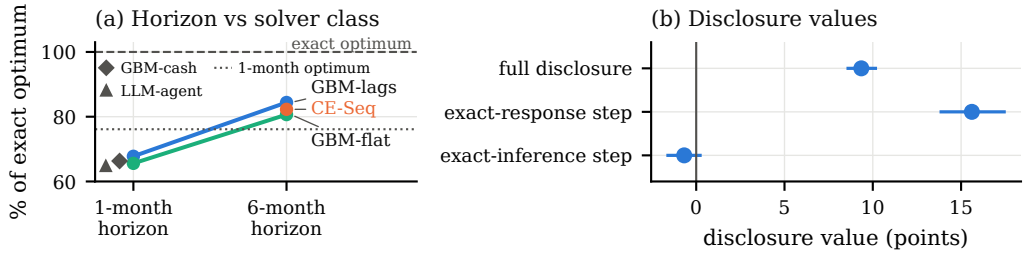

Figure 2: Closed-loop performance on BASE. (a) Value at the two horizons. (b) Disclosure values with 95% intervals.

Full disclosure to GBM-lags, the exact response surface and exact latent inference together, recovers 9.35 of its 15.6 points of distance to the exact optimum (Fig. 2b). Within the pre-registered ladder behind that disclosure (Appendix D), handing GBM-lags exact latent inference is worth nothing measurable, and the dominant step replaces its fitted response with the exact one. That step exceeds the total because the ladder's other steps are negative.

Five theory-motivated refinements of GBM-lags were evaluated under identical data and protocol, and none outperforms GBM-lags itself (Appendix D). An oracle-side diagnostic bounds what those estimators were missing. A linear model reading the last six months of readings and actions recovers the oracle's hidden score with R^2 of 0.56 per contact channel (0.06 when a single linear model is fit across the three channels), so the unextracted information is present in the logged observables and is linearly recoverable. The logged worlds explore with $\varepsilon = 0.03$ and never explore the no-contact action, so the constraint is one of identification, and under that design none of the estimator classes recovers it.

Certainty-equivalent chaining inflates a planner's self-reported value, the value it computes for its own plan, by approximately 11% of that plan's value under the exact belief, even when the planner's belief is exact. The measurement is on BASE (4,000 loans, 512 draws from the belief, one geometry, no interval available). Graded and self-reported values are reported separately throughout.

3.2 MEMORY

A single transition parameter δ relocates how much of the outcome-relevant information is kept in a client's history (Fig. 3a). The sweep runs on pilot worlds, smaller unreleased worlds of 100,000 borrowers over 12 months generated on the members' geometry, and members are released at $\delta \in \{0, 0.5, 4\}$. At $\delta = 0.5$ the separation between the history probe and the snapshot probe rises from $+0.114$ to $+0.177$ while the snapshot score moves by less than half a point, and at $\delta = 4$ the headroom above the history probe opens to $+0.416$ as the snapshot score falls, a fitted contrast distinct from the no-fit margin of Eq. 4 reported below. Across five independent draws of the world's matrices (Fig. 3b), each pre-registered ordering clause holds in four of five draws, the snapshot score falls at $\delta = 4$ in five of five, and the paired within-draw separation is positive in all fifteen draw–δ cells. The ordering replicates, the magnitude at each δ is draw-specific, and each member is characterized individually.

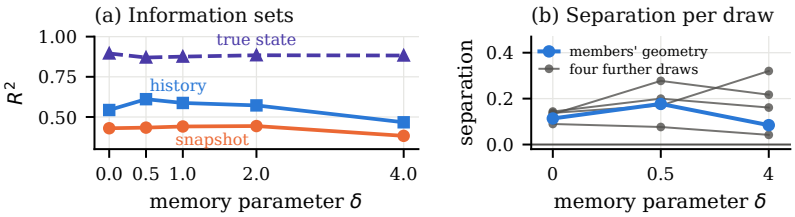

Figure 3: The memory axis. (a) Scores of the three information sets against δ. (b) Separation across five geometry draws.

Each recurrent configuration, a gated recurrent unit (GRU) or long short-term memory (LSTM) network, reads a loan's full history of readings and actions and compresses it into a vector that, with the current month's snapshot, feeds a small network predicting the probes' target, so its out-of-fold R^2 minus the history probe's is its memory value above the history probe. On MEM-4, where the information bound lies 0.289 in R^2 above the history probe, the recurrent configurations of Table 6 collect between approximately zero and 37% of that margin at frozen standard training budgets, while the exact filter collects 69% (Fig. 4a, Appendix F). The bound counts only the belief's error as unexplained, while the hidden traits and the outcome noise are also out of reach, so no model attains it and these shares are conservative. The three larger configurations reached the epoch cap in every fold, still improving in most folds of the two large ones and in about half of the long-budget one's, so the shares are budget-capped lower bounds. The spread across configurations reflects capacity and optimization budget. The ordering among them at a single budget is not a finding of this paper. At $\delta \leq 0.5$ every recurrent configuration ties within two points of R^2.

3.3 PROCESS NOISE

The re-observation gain is the value of re-planning every month on the updated belief minus the value of executing the plan chosen at the decision point. On BASE it is approximately zero even for the exact filter, and under process noise it is positive and increases with σ (Fig. 4b). Re-planning with the exact belief gains 0.57% of the member's upper reference at $\sigma = 0.1$ and 3.7% at $\sigma = 0.2$. In expected recovered value per loan, in the world's currency units, the gains are $+9.4$ ($[4.2, 14.9]$) and $+87.1$ ($[62.9, 116.1]$). Intervals condition on the 24 shared innovation paths.

3.4 REAL DATA

Requiring no outcome-level ground truth, the outcome-level separation transports to real data. On 20,677 public Fannie Mae delinquency spells (Fannie Mae, 2024) (episodes of delinquency), adding

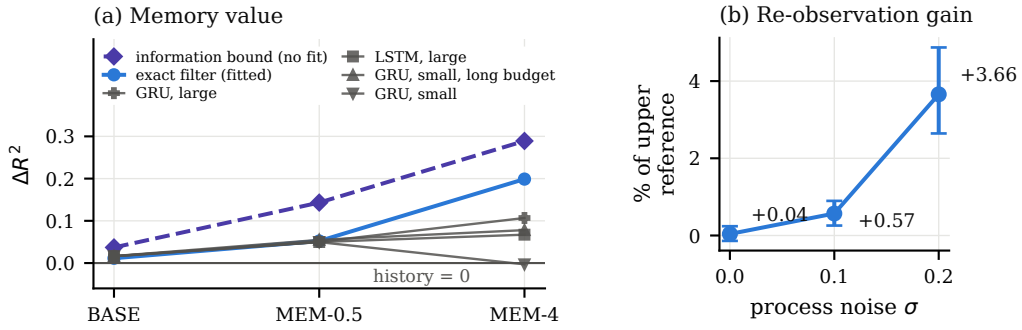

Figure 4: Memory and re-observation. (a) Memory value above the history probe on the three members without process noise. (b) Re-observation gain against process noise, with 95% intervals.

the pre-entry block (twelve months of delinquency states, prior-spell count, months since the previous cure) raises the area under the receiver operating characteristic curve (AUC) for cure prediction, cure meaning a return to current status within six months, by $+0.041$ ($[0.031, 0.049]$, grouped bootstrap by loan, 1,000 resamples). The same statistic on the released members spans $+0.006$ to $+0.010$, all three intervals excluding zero (per-member values in Appendix I). The real portfolio sits well above the span of the released worlds, so the synthetic members understate outcome-level memory value relative to it. The outcome-level metric compresses history effects relative to the information probes of Section 3.2, and across the members it decreases slightly with δ (Appendix I). The real side is a mortgage portfolio observed without contact actions, its label is the six-month cure against the synthetic recovery within nine months, a disclosed mismatch, and each side was computed exactly once under a pre-registered protocol.

4 Discussion

On BASE, the horizon contrast within a solver class is several times any contrast between solvers at a fixed horizon, and full disclosure recovers most of the gap of GBM-lags, while handing it exact latent inference is worth nothing measurable. Because none of five theory-motivated refinements outperforms GBM-lags while the missing information is linearly recoverable from the last six months of readings and actions, the binding constraint is the exploration design itself. Whether a different logging policy recovers the information is an experiment the generator supports and this paper does not run. The worth of re-planning on fresh observations is a measurable property of each world, and because self-reported values inflate even at exact inference, external grading is required to compare planners.

References computable from the world's equations require a controlled world. The released worlds instantiate one domain. Their dynamics are conditionally tractable by design, and the architecture results are budget-capped lower bounds. The reported magnitudes are measured on the members' single geometry, and across the four further draws only their orderings replicate. Interval widths reflect the chosen 4,000-loan grading cohort, and the recurrent configurations carry two training seeds outside MEM-4. On real data, the transport comparison carries a disclosed label mismatch, cure within six months against recovery within nine, with one run per side. The transport statistic can serve as a calibration target for the generator, though outcome-level effects compress, so calibration must move parameters beyond the memory parameter alone. Sequence models trained across many generated worlds could also be evaluated for in-context inference of a member's dynamics, a study the single configuration of Appendix G does not attempt.

5 Related work

Existing evaluation practice grades solvers by final performance. Offline reinforcement learning (Levine et al., 2020) suites such as D4RL (Fu et al., 2020) standardize datasets and report policy returns from logs more exploratory than deployed policies produce (Qin et al., 2022). Off-policy

evaluation benchmarks, COBS (Voloshin et al., 2021) and DOPE (Fu et al., 2021), compare value estimators against returns from running the policy. Identification from logs rests on the randomization the logging design provides (Bottou et al., 2013), and unobserved confounding yields only bounds, which widen with the horizon (Namkoong et al., 2020). Reinforcement learning has been deployed for collections (Abe et al., 2010), evaluated by importance weighting on logged outcomes with no reference optimum. These suites rank methods. The optimum, the value of information (Howard, 1966), and the price of each disclosed ingredient are not among their reported objects.

Evaluation practice in model-based reinforcement learning judges a learned model by the return of the policy trained against it (Janner et al., 2019; Yu et al., 2020), and the value-equivalence line (Grimm et al., 2020) judges models by the value-based planning they support, a position DECODE operationalizes. The memory axis stresses the memory component of the world-model decomposition (Ha & Schmidhuber, 2018), isolating sequence encoders behind a shared readout. Learned recurrent state-space models are graded by downstream return (Hafner et al., 2021), with no reference belief to compare against. Planning under the released physics is a partially observable Markov decision problem (Kaelbling et al., 1998). Value overestimation is classical, from the maximization step (Thrun & Schwartz, 1993; van Hasselt et al., 2016), from selection among noisy estimates (Smith & Winkler, 2006), and from selecting a policy on the sample that estimated its model (Mannor et al., 2007). The inflation measured here is a Jensen bias of certainty-equivalent chaining, read against the executed plan's value under the exact belief.

Diagnostic suites with tractable environments make different trades. bsuite (Osband et al., 2020) scores regret normalized between random and optimal policies, one number per experiment, decomposing no single gap. MDP Playground (Rajan et al., 2023) varies hardness dimensions over toy environments with known optima and rescales end performance so the optimum reads 100, which uses the optimum only as a normalizing constant. POPGym (Morad et al., 2023) isolates memory models on partially observable tasks and normalizes by the best observed trial, so no optimum enters its scores, and Memory Gym and Memory Maze (Pleines et al., 2023; Pasukonis et al., 2023) report completion rates or a privileged-information oracle score. None prices what memory adds against a reference computed from the environment's equations without fitting, where DECODE prices it in R^2 against a no-fit analytic bound and in value against a certainty-equivalent filtering reference. Which architecture wins depends on the task's memory length (Ni et al., 2023), and recurrent baselines are strong once implementation is controlled (Ni et al., 2022). DECODE reports the contrast conditional on a memory parameter and against a no-fit bound. A diagnostic study of deep Q-learning decomposes the gap to exact tabular solutions into estimator-internal error sources, characterizing one algorithm class on a fixed domain set (Fu et al., 2019), and its oracle-stopping headroom is distinct from the self-report inflation measured here. Contemporaneous work generates environments with certified optima along a difficulty ladder (Ibrahim et al., 2026), reporting one optimality gap in a fully observed setting with no inference or memory component. DECODE combines what these keep apart, a reference recomputed at every rung of a hardening ladder and a same-task gap decomposed against it into informational ingredients, defined by what the world reveals, independently of any one estimator's internals.

6 Conclusion

DECODE turns the gap components the instrument isolates into measured objects. On BASE, the within-class horizon contrast is worth 15 to 17 points where no fixed-horizon contrast between solver classes exceeds 4. Full disclosure recovers nine of the fifteen points GBM-lags lacks, while handing it exact latent inference is worth nothing measurable, and a channel-by-channel diagnostic locates the binding constraint in the exploration design, which none of the tested estimator classes removes. Where memory physics is shallow, every recurrent configuration ties, so results on such worlds cannot rank memory architectures. On BASE, certainty-equivalent chaining inflates a planner's self-report even at the optimal-inference limit, so comparing planning systems requires external grading. References of this kind require a controlled world, and the released worlds' scope follows, one domain, dynamics conditionally tractable by design, architecture measurements that are budget-capped lower bounds. Within that scope, the measurements bound what effort can buy.

ETHICS STATEMENT

The paper's object is an evaluation instrument, and its domain instantiation is consumer debt collection, a setting with direct borrower-welfare implications. Three considerations apply. First, all optimization results in this paper are measured on synthetic worlds. No policy was applied to any real borrower, and the paper makes no deployment claim. The single real dataset is a public, de-identified loan-performance research file used only to compute an aggregate predictive statistic, and no individual-level inference is drawn or attempted. Second, the instrument's design internalizes part of the borrower's burden by pricing contact through per-channel costs and an explicit contact-fatigue state, and the released acceptance checks constrain aggregate recovery behavior to realistic bands. These are modeling choices, and a deployed system would require welfare constraints that the instrument does not itself impose. Third, the measurement direction of this work supports oversight of automated decision systems. External grading against references the operator does not control, and the finding that planners inflate their self-reported value, are tools for auditing decision systems of this kind, and the benchmark's controlled disclosures are designed to make such audits reproducible.

REPRODUCIBILITY STATEMENT

The supplementary material contains an anonymized reproduction package. It holds the world generator with the released member configurations and their seeds, the acceptance-check reports and regeneration manifests for all five members (including the re-anchoring audit trails and the amended clause record), the closed-loop harness with the references and the probes, the code of every graded solver, the per-fold stopping logs of the capped recurrent and attention configurations, the LLM-agent's frozen policy file, the per-loan values and executed plans behind every closed-loop interval on BASE, and the figure-generation code. CE-Seq ships with its trained weights and its training code, GBM-cash with its fitted model, and GBM-lags and GBM-flat are refit from the regenerated training world. The Monte Carlo intervals of the noisy members are regenerated from the recorded seeds by the included script. Regeneration of any member requires the manifest's configuration hash, seeds, and pinned generation batch size, all recorded in the package. The real-data experiment uses the public Fannie Mae single-family loan-performance file (Fannie Mae, 2024), freely available upon registration. The data licence does not allow redistribution, so the package records the analysed loans by a fingerprint (their number, their count per quarterly file, and a hash of the identifier list) instead of listing them. It also holds the spell-construction and evaluation code and the frozen protocol. The identifier list is available from the authors to holders of the data licence, and with it the reported statistic is reproducible from the public source.

AI USE STATEMENT

One evaluated solver is itself a language-model-authored decision rule and is an object of study. Its construction and grading protocol are described in Appendix E, and the authoring model was Claude Fable 5 (Anthropic). Beyond that policy, language models were used in the research and in the preparation of this paper as follows:

Research

The world generator and the evaluation harness were implemented with language-model assistance under human review, and the released synthetic worlds are seeded, deterministic outputs of that code, generated from equations and acceptance checks specified and approved by the authors. The instrument's conceptual design and the pre-registered experimental protocols were developed in an author-directed loop with language-model assistance, with each decision recorded and approved by the authors before any run.

Synthetic data

Every released world is synthetic. The generating code, the member configurations, and the acceptance checks were written with language-model assistance and approved by the authors. The data themselves are deterministic outputs of that code under recorded seeds, and no language model produced, edited, or selected any record.

Writing

Prose was reviewed by a language model from a ledger of claims approved by the authors. Automated adversarial review agents checked claims, numbers, and citations, recording verbatim supporting quotes with their loci for every reference, and the resulting corrections were applied under the authors' review. Candidate related work was retrieved with language-model assistance, and each reference was verified against the source text before citation. Figure-generation code was written with language-model assistance.

Tasks without AI involvement

No experimental result reported here was produced by a language model, apart from the evaluated LLM-agent above. Approval of every claim, protocol, and release decision was made by the authors, and no language model authored the acceptance criteria after seeing generated data.

Verification

All AI-assisted work was reviewed by the authors. Research code carries a test suite. The grading harness includes a consistency check under which the frozen and closed-loop evaluations of the rule-based policy coincide on BASE (a shared-path check, not an end-to-end validation). Citations were verified by type against the source texts with verbatim quotes, and independent audits recomputed the reported numbers from the frozen artifacts. The authors take full responsibility for the content of this paper. In the categories of the submission form, the uses above are aiding or polishing writing, retrieval and discovery, research ideation or execution, drafting sections of the paper, and generating synthetic datasets. No language model was used to prove mathematical claims.

REFERENCES

Naoki Abe, Prem Melville, Cezar Pendus, Chandan K. Reddy, David L. Jensen, Vince P. Thomas, James J. Bennett, Gary F. Anderson, Brent R. Cooley, Melissa Kowalczyk, Mark Domick, and Timothy Gardinier. Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 75–84, 2010.

Yaakov Bar-Shalom and Edison Tse. Dual effect, certainty equivalence, and separation in stochastic control. *IEEE Transactions on Automatic Control*, 19(5):494–500, 1974.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14: 3207–3260, 2013.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of EMNLP*, 2014.

David R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

Fannie Mae. Single-family loan performance data. `https://capitalmarkets.fanniemae.com/credit-risk-transfer/single-family-credit-risk-transfer/fannie-mae-single-family-loan-performance-data`, 2024.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

Justin Fu, Aviral Kumar, Matthew Soh, and Sergey Levine. Diagnosing bottlenecks in deep Q-learning algorithms. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2021–2030, 2019.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R. Zhang, Yutian Chen, Aviral Kumar, Cosmin Paduraru, Sergey Levine, and Tom Le Paine. Benchmarks for deep off-policy evaluation. In *International Conference on Learning Representations*, 2021.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with discrete world models. In *International Conference on Learning Representations*, 2021.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966.

Sinan Ibrahim, Grégoire Ouerdane, Hadi Salloum, Henni Ouerdane, Stefan Streif, and Pavel Osinenko. Benchmarking reinforcement learning via stochastic converse optimality: Generating systems with known optimal policies. *arXiv preprint arXiv:2603.17631*, 2026.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, 2019.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

Akshay Krishnamurthy, Keegan Harris, Dylan J. Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context? In *Advances in Neural Information Processing Systems*, 2024.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Shie Mannor, Duncan Simester, Peng Sun, and John N. Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

Steven Morad, Ryan Kortvelesy, Matteo Bettini, Stephan Liwicki, and Amanda Prorok. POP-Gym: Benchmarking partially observable reinforcement learning. In *International Conference on Learning Representations*, 2023.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. In *Advances in Neural Information Processing Systems*, 2020.

Tianwei Ni, Benjamin Eysenbach, and Ruslan Salakhutdinov. Recurrent model-free RL can be a strong baseline for many POMDPs. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16691–16723, 2022.

Tianwei Ni, Michel Ma, Benjamin Eysenbach, and Pierre-Luc Bacon. When do transformers shine in RL? decoupling memory from credit assignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50429–50452, 2023.

Ian Osband, Yotam Doron, Matteo Hessel, John Aslanides, Eren Sezener, Andre Saraiva, Katrina McKinney, Tor Lattimore, Csaba Szepesvari, Satinder Singh, Benjamin Van Roy, Richard Sutton, David Silver, and Hado van Hasselt. Behaviour suite for reinforcement learning. In *International Conference on Learning Representations*, 2020.

Jurgis Pasukonis, Timothy Lillicrap, and Danijar Hafner. Evaluating long-term memory in 3D mazes. In *International Conference on Learning Representations*, 2023.

Marco Pleines, Matthias Pallasch, Frank Zimmer, and Mike Preuss. Memory gym: Partially observable challenges to memory-based agents. In *International Conference on Learning Representations*, 2023.

Rong-Jun Qin, Xingyuan Zhang, Songyi Gao, Xiong-Hui Chen, Zewen Li, Weinan Zhang, and Yang Yu. NeoRL: A near real-world benchmark for offline reinforcement learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, pp. 24753–24765, 2022.

Raghu Rajan, Jessica Lizeth Borja Diaz, Suresh Guttikonda, Fabio Ferreira, André Biedenkapp, Jan Ole von Hartz, and Frank Hutter. MDP playground: An analysis and debug testbed for reinforcement learning. *Journal of Artificial Intelligence Research*, 77:821–890, 2023.

James E. Smith and Robert L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

Sebastian Thrun and Anton Schwartz. Issues in using function approximation for reinforcement learning. In Michael Mozer, Paul Smolensky, David Touretzky, Jeffrey Elman, and Andreas Weigend (eds.), *Proceedings of the 1993 Connectionist Models Summer School*, pp. 255–263. Lawrence Erlbaum, Hillsdale, NJ, 1993.

Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double Q-learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 2094–2100, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Cameron Voloshin, Hoang M. Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, 2020.

A GLOSSARY

The terms defined in the body of the paper are collected here for reference, in alphabetical order (Table 2).

Table 2: Glossary.

Term	Meaning
BASE, MEM-0.5, MEM-4, NOISE-0.1, NOISE-0.2	the released members, BASE with $\delta = \sigma = 0$, the MEM members with $\delta = 0.5$ and 4 and $\sigma = 0$, the NOISE members with $\sigma = 0.1$ and 0.2 and $\delta = 0$
belief	the probability distribution of the hidden state given everything a solver has observed
certainty-equivalent (CE)	planning on point predictions as if they were certain, ignoring their uncertainty
closed-loop grading	grading in which each solver re-decides every month on the readings its own actions generate
decision point	month six of an evaluation loan, when a solver takes over after six months of public history, with six months left to decide
disclosure	the replacement of one input of a graded solver by its ground truth
exact filter	the closed-form belief over the hidden state given a client's readings and actions
exact optimum	the value of the best plan computed with the true hidden state at the decision point, available on members without process noise

Term	Meaning
exploration rows	logged loan-months on which the logging policy drew a contact channel and a discount at random
gap	the difference between the exact optimum and a solver's value, in points
geometry	one random draw of the fixed matrices and vectors in a world's equations. The five released members share one geometry
headroom	the true-state probe score minus the history probe score
hidden score	the part of the repayment logit that depends on the hidden state through the acting channel, $\kappa\, u_c^\top z_t$
hidden traits	four fixed borrower characteristics that enter repayment and are never observed
information bound	the ceiling, computed without fitting, on the R^2 of any probe that learns about the hidden state only from readings and actions
information clauses	pre-registered floors and ceilings on probe scores, and on BASE also on values, that a member with an exact reference must pass
logging policy	the fixed rule that generated the training logs
member	one released setting of the memory parameter and the process-noise scale
memory parameter δ	the scale of the rotation each action applies to the hidden state, which makes actions interact with the state
myopic solver	a solver that chooses each month the action with the highest expected value for that month alone
one-month optimum	the best policy that plans one month at a time with the true hidden state
oracle	the grader's side of the world, which holds the hidden state and the equations
pilot world	a smaller unreleased world, of 100,000 borrowers over 12 months, used for sweeps and construction checks
point	one percentage point of a member's upper reference
probe	the out-of-fold R^2 of a gradient-boosted model predicting the true repayment logit from one information set, snapshot, history, or true state (Table 1)
process-noise scale σ	the standard deviation of the monthly innovation in the hidden state
reading	the noisy two-dimensional linear observation of the hidden state
realism bands	ranges into which the share of loans recovered within one month and within nine months must fall
re-anchoring	regenerating a member with an adjusted base rate after its aggregate recovery fell outside the realism band
re-observation gain	the value of re-planning every month on the updated belief minus the value of executing the plan chosen at the decision point
reference	a value computed from the world's equations, against which a solver's result is read
reference ladder	the set of references, whose form follows the physics, closed form, analytic, or Monte Carlo with certified precision
reference solver	a labeled copy of a graded solver that receives a disclosure and never competes
rotation	the orthogonal matrix $Q_w = \exp(w\,\delta\,S)$ through which an action turns the hidden state
self-reported value	the value a planner computes for its own plan
separation	the history probe score minus the snapshot probe score, and at outcome level the gain in AUC from adding history
sequential solver	a solver that plans the whole remaining horizon every month, executes the first action, and re-plans
solver	a decision system graded by the instrument, built from the training logs
solver class	a solver's architecture and estimator, which can be graded at more than one horizon
upper reference	the normalizer of each member, the exact optimum without process noise and the value of monthly re-planning with the true state under noise

B MEMBERS

The five members of Section 2 are listed in Table 3 with their transition physics, their parameters, and the outcome of their acceptance checks. The information clauses set floors on the separation or the headroom of the probes of Table 1, so that the history or the hidden state carries information the current month lacks, and, on the memory members, ceilings on what the current month alone predicts, so that the reading does not leak the hidden state. On BASE the clauses are also stated on values, requiring that planning over the horizon matter and that an engine without history stay well below the optimum. The first requires the exact optimum to exceed the one-month optimum by

15 to 40 points, and BASE measured 27.1 on a 100,000-borrower prefix of its training world. The second caps a history-free engine at 85 points, and BASE measured 59.2 there, before any solver of Section 3 was graded. The two memory members also pass a per-loan consistency check, under which the plan optimum is never below the one-month optimum, and on BASE the same holds on every graded loan. The thresholds of every clause are recorded with the release.

Each release carries the machine-readable check report and a manifest recording the configuration hash, the seeds, and the pinned generation batch size that regeneration requires. Check criteria were fixed before generation, with one exception. An information clause on MEM-0.5 failed and was replaced once before release by a clause targeting the same mechanism, and the amendment is in the release records. Three members were re-anchored, that is, regenerated with an adjusted base rate after their aggregate recovery fell outside the realism band, two on pilot or check worlds before definitive generation and one after its definitive world failed the check. The release records each adjustment with its audit trail.

Two further verifications were run at construction time on pilot worlds. With $\delta = 0$ the generator reproduces bit for bit the output of its version without the memory parameter, so the parameter changes nothing when it is off, and the grader's replay of each client's hidden-state path under the logged actions agrees with the generator to within 10^{-12}.

Table 3: Released members. All pass realism bands on recovery rates, and the three $\sigma = 0$ rows additionally pass pre-registered information clauses. Three members were re-anchored in base rate after a realism-band failure, and one member's clause set was amended once, both marked in the table.

Member	Transition physics	Memory δ	Noise σ	Checks
BASE	linear, deterministic	0	0	pass, re-anchored
MEM-0.5	action-rotation	0.5	0	pass, one clause amended
MEM-4	action-rotation	4	0	pass, re-anchored
NOISE-0.1	linear, stochastic	0	0.1	pass
NOISE-0.2	linear, stochastic	0	0.2	pass, re-anchored

C SOLVERS

Each graded solver of Section 2.5 is specified in Table 4 by the information it reads, the model it fits with the rows it is trained on, its decision rule, and the horizons at which it was graded.

D LADDER

The full disclosure of Section 3.1 is reached from GBM-lags in four pre-registered single-factor steps, each of which changes one input and is graded on the same 4,000 loans with one shared-index bootstrap (Table 5). The three steps that make the engine better informed about the state all lower its value, and only the exact response raises it, which is why that step exceeds the total. The pattern is consistent with a fitted response surface being reliable only on the support of its training data. Rolling the belief through the imagined actions of the search carries the engine into states the logging policy never produced, and there its fitted response extrapolates. The remaining 6.25 points between the full disclosure's 93.75 and the exact optimum are what planning on the true hidden state adds to planning on the exact belief. The intervals of the ladder contrasts divide each bootstrap resample's mean difference by that resample's mean optimum.

Each measured directly against GBM-lags, the three intermediate engines of the ladder score -0.75 ($[-1.18, -0.34]$), -1.43 ($[-2.37, -0.55]$) and -6.26 ($[-7.77, -4.80]$). Two further refinements complete the five of Section 3.1. A structured estimator replaces the history features of GBM-lags with a linear correction in logit space, fit on top of a history-free GBM, which matches the form of the true response, and it scores -0.98 ($[-1.86, -0.12]$) against its pre-registered comparator, the engine with exact fatigue. An 11-lag window, the one variant that adds public information, is indistinguishable from GBM-lags at $+0.63$ ($[-0.12, +1.47]$). The pattern is consistent with the

Table 4: Graded solvers. Horizons are in months, and a solver graded at both horizons forms a within-class horizon pair.

Solver	Reads	Model and training rows	Decision rule	Horizons
GBM-lags	public state, current reading, last six months of readings and actions	two calibrated gradient-boosted classifiers, one for contact months fit on exploration rows and one for no-contact months fit on logged no-contact months, training-world months six onward	myopic, or the exact search with its features held at the decision month	1, 6
GBM-flat	public state, current reading	as GBM-lags	as GBM-lags	1, 6
CE-Seq	the account's full observed history, through a learned belief	recurrent belief model with contact and no-contact heads, each with its own isotonic calibrator, full training panel	exact search chaining predicted probabilities, future readings masked	6
GBM-cash	public state, current reading	gradient-boosted model of expected monthly cash with a Tweedie loss, whole logged book, temporal validation split	myopic	1
LLM-agent	delinquency ageing and outstanding balance	state-conditional rules written once from a censored copy of the training logs	rules applied each month	1

Table 5: The single-factor ladder from GBM-lags to full disclosure on BASE, in % of the exact optimum. Each row changes one input of the row above, and the steps sum to the full disclosure by identity.

Engine	Change from the row above	Value	Step
GBM-lags	none	84.40	
exact fatigue	exact contact-fatigue state instead of the engine's smoothed fatigue feature	83.65	-0.75
exact inference	exact belief over the hidden state instead of the six months of history	82.97	-0.68
rolled belief	exact belief propagated through the search instead of frozen at the decision month	78.14	-4.83
exact response	true response surface instead of the fitted one	93.75	$+15.61$
full disclosure	all four changes together		$+9.35$

free-form fit of GBM-lags absorbing interactions between the state, the history, and fatigue that each decomposition freezes out.

A separate experiment decomposes the response disclosure under its own protocol, with spans not additive to the ladder. Per-loan discrimination of the contact response accounts for $+6.7$ points ($[+5.7, +7.8]$), effect propagation across months for $+3.3$ ($[+2.3, +4.1]$), and the organic component, the response on no-contact months, is slightly negative at -0.32 ($[-0.63, -0.05]$). The negative organic reading is consistent with the logging design, whose never-randomized no-contact action leaves the organic response of GBM-lags fit on confounded rows.

E LLM-AGENT

LLM-agent is a fixed decision rule authored by a large language model before evaluation. The model investigated a censored copy of the training world's logged decision table, in which the per-row randomization labels were published. Its access was read-only queries over the row-level records, supplemented by randomized-subset and observational effect estimates with standard errors and per-segment overlap summaries, within a budget of eight interactive rounds, with no holdout within that table and without access to the evaluation world. Columns encoding outcome history were withheld from the investigation because no payment is realized inside the closed-loop protocol, so a rule con-

ditioned on them has no defined counterfactual value. Columns encoding post-action status were withheld because they encode the month's outcome label, and 36 state columns remained available. From this investigation the model wrote seven state-conditional rules and an unconditional default that makes the policy total. The rules read two public fields, delinquency ageing and outstanding balance, and each names a channel and a settlement offer. Above one year of delinquency ageing the logged design carries no randomized coverage, and the two rules covering that band rest on observational contrasts, a restriction recorded with the policy. Language models do not explore robustly in bandits from raw in-context histories, the one successful configuration relying on an externally summarized history (Krishnamurthy et al., 2024). The policy here is likewise authored from summarized statistics of a logged table, as a fixed rule that does not explore in context. The rule set was frozen before evaluation and applied month by month. Because the evaluated member is deterministic and the harness serves the action history, both fields roll deterministically between observations, and the policy's closed-loop and frozen-information evaluations therefore coincide exactly, a consistency check that BASE makes exact by construction. It verifies that the two evaluation paths share one implementation and cannot detect errors common to both. Graded on BASE, the policy reached 64.74% of the exact optimum in a single observation dated August 2026, below every fitted solver (Table 8). The authoring model is named in the AI use statement.

F MEMORY VALUES

The memory values of Fig. 4a on MEM-4 are tabulated in Table 6, with seed dispersions, the share of the margin each configuration collects, and whether it reached the epoch cap.

On BASE, the information bound behaves as a ceiling. A fitted pipeline on the exact filter's outputs reaches R^2 of 0.649 against a 0.674 information bound, a fitted recurrent memory (0.653) lands between the two, and the history probe is below both at 0.637. These levels are from the training world and are not comparable to the pilot-world sweep of Fig. 3.

Table 6: Memory value on MEM-4 above the history probe (mean $\pm$ population sd over five training seeds) and the share each collects of the 0.289 margin between the history probe and the information bound. Recurrent encoders are LSTM (Hochreiter & Schmidhuber, 1997) and GRU (Cho et al., 2014) variants. Their values are budget-capped lower bounds where the logs show the cap was reached, and the ordering at a single budget is not a finding.

Configuration	ΔR^2 vs history	share	at epoch cap
recurrent (GRU), small	-0.003 ± 0.003	≈ 0 (tie)	no stopping log
recurrent (GRU), small, long budget	$+0.078 \pm 0.002$	27.0%	yes
recurrent (LSTM), large	$+0.067 \pm 0.006$	23.2%	yes
recurrent (GRU), large	$+0.107 \pm 0.002$	36.9%	yes
exact filter (fitted)	$+0.199$	68.8%	
information bound (no fit)	$+0.289$	100%	

G ATTENTION

Alongside the recurrent configurations of Table 6, one attention-based sequence model was evaluated with the same readout head, folds, and data. The encoder is a two-layer transformer (Vaswani et al., 2017) of model width 64 with four heads, causal masking, and learned positions. Its epoch cap and patience match the standard recurrent configurations' at a lower learning rate, and Table 7 states the budgets next to the measured memory values. One recurrent configuration of Table 6 ran a longer budget, so shares credited to it and the values reported here are not comparable at equal optimization effort. Each member cell trains three folds per seed. Every transformer fold reached the epoch cap, twelve of the fifteen MEM-4 folds were still improving there, and three had peaked before the cap, so the MEM-4 value is, like the recurrent shares of Table 6, a budget-capped lower bound. A single small configuration at one budget supports no conclusion about attention-based memories, and the per-fold stopping logs are recorded with the measurement.

Table 7: The attention-based configuration. Top, its training budget beside the recurrent configurations of Table 6. Bottom, its memory value above the history probe per member. The interval is the population standard deviation over seed means, per-seed values are shown where two seeds were run, and on MEM-0.5 the sign is not stable across the two seeds.

configuration	epoch cap	patience	learning rate
attention (this appendix)	30	5	3×10^{-4}
recurrent (Table 6)	30	5	10^{-3}
recurrent, long budget (Table 6)	100	15	10^{-3}
member	**seeds**	**memory value**	**per-seed**
BASE	2	-0.018	-0.018, -0.018
MEM-0.5	2	-0.001	-0.009, $+0.007$
MEM-4	5	-0.089 ± 0.004	

H CLOSED-LOOP VALUES

Every closed-loop value on BASE, including the two reference bounds, a random policy, and the single-observation LLM-agent reported in Section 3.1, is listed in Table 8.

Table 8: Closed-loop values on BASE, in % of the exact optimum (paired bootstrap protocol of Section 3). Every solver row is plotted in Fig. 2(a), and the random policy sets the floor.

Solver	1-month horizon	6-month horizon
GBM-lags	67.78	84.40
GBM-flat	65.59	80.65
CE-Seq		82.28
GBM-cash	66.32	
LLM-agent	64.74	
random policy	29.11	
1-month optimum	76.13	
exact optimum		100

I TRANSPORT

The per-member outcome-level separation behind Section 3.4 is given, with its intervals, in Table 9.

Table 9: Outcome-level separation (out-of-fold ΔAUC of adding pre-entry history over five loan-grouped folds, with grouped-bootstrap intervals) computed with the identical estimator and fold policy on the real portfolio and on the released members. The real-side outcome is cure within six months and the synthetic label is recovery within nine, a disclosed mismatch, so the cross-side comparison is qualitative.

Data	ΔAUC	95% CI
Fannie Mae spells (real, $n = 20{,}677$)	$+0.041$	$[0.031, 0.049]$
BASE	$+0.010$	$[0.007, 0.013]$
MEM-0.5	$+0.007$	$[0.004, 0.009]$
MEM-4	$+0.006$	$[0.003, 0.009]$