# CLOSING THE LOOP: AUDITABLE HALF-LIFE PREDICTION WITH OPEN-SPACE RETRIEVAL AND CAPACITY-AWARE SCHEDULING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Spaced repetition requires accurate memory-state prediction and effective scheduling under limited review capacity. We present a closed-loop framework that couples Open-Space Retrieval for Half-Life Regression (OSR-HLR) with the OSR Adaptive Calibration-Regret Index (OSR-ACRI). OSR-HLR combines relative half-life estimation with open-space retrieval, which transfers deterministic, source-traceable residual evidence from recurring and related historical memory states, while OSR-ACRI converts half-life estimates into capacity-aware review decisions using online calibration, marginal delay cost, and queue age. Across MaiMemo, Anki, and EdNet-Recall, under source-only protocols that keep every Test outcome out of training and retrieval, OSR-HLR achieves the lowest H-SMAPE among 27 baselines in all 15 dataset–protocol settings. It reduces H-SMAPE by **17.40%**, **2.59%**, and **42.46%**, respectively, under the Row protocol, and by **11.36%** on held-out Anki users. Under a fixed scheduler, improved predictions reduce 365-day regret by **8.7%–15.8%**. With the predictor fixed and workload matched, OSR-ACRI improves retention AUC by **11.1%**, reduces time below target by **7.5%**, and improves final recall by **19.8%**. Together, the two components provide an auditable prediction-to-scheduling loop for efficient spaced repetition.

## 1 INTRODUCTION

A standard forgetting-curve model gives the probability of recall after an elapsed time $\Delta t$ as $R(\Delta t) = 2^{-\Delta t/h}$, where the half-life $h$ is the delay at which recall probability drops to $0.5$ (Ebbinghaus, 1913). Under this model, the half-life is more than a regression target: it directly sets when an item should next be reviewed. An overestimated $\hat{h}$ postpones review past the point of reliable recall, so the item is more likely to be forgotten. An underestimated $\hat{h}$ triggers early reviews that use learner time but add little retention. Accurate half-life estimates are necessary but not sufficient. When more items are due than the daily review budget allows, the scheduler must prioritize the items whose recall would suffer most from one more day of delay, and it must also guarantee that no item is deferred indefinitely. Half-life prediction and capacity-aware scheduling are thus distinct but coupled problems.

Existing approaches to recall prediction range from compact cognitive update rules, such as HLR (Settles & Meeder, 2016), DASH (Lindsey et al., 2014), ACT-R (Anderson et al., 1997), DHP (Ye et al., 2022), and DSR-style stability updates (Su et al., 2023), to recurrent and knowledge-tracing models (Piech et al., 2015; Nagatani et al., 2019; Ma et al., 2023). These models encode useful inductive biases, but two practical difficulties remain. First, observed recall depends on both state-level half-life and the query interval. The same memory state can occur at multiple elapsed times, making direct half-life regression sensitive to multiplicative scale variation and requiring the predictor to separate state strength from query timing. Second, recurring and structurally related review histories can provide historical evidence about systematic prediction errors. The outcome-conditioned patterns in Figure 1 motivate preserving outcome structure in these state relations. Such evidence must come only from previously observed states and outcomes, never from the review being predicted. Together, these difficulties raise a central question: can half-life prediction ❶ separate state strength from query-time scale, ❷ reuse evidence from recurring and related historical states under an explicit visibility boundary, ❸ generalize across disjoint learners, items, and states, and ❹ translate the resulting estimates into effective review decisions under a capacity constraint?



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.

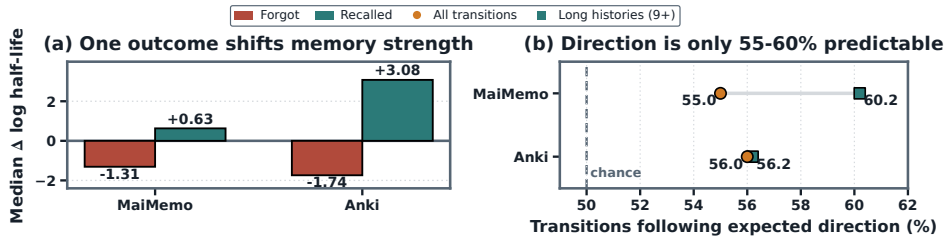


Figure 1: Outcome-conditioned changes in memory strength across two spaced-repetition system (SRS) review logs. Outcome-conditioned differences vary by dataset, motivating memory-state representations that preserve outcome structure when estimating and transferring evidence.

OSR-HLR addresses this problem through a three-stage prediction process. First, a base estimator predicts half-life relative to the query interval, reducing multiplicative variation in the prediction target. Second, observations sharing the same memory state are reconciled into a state-level estimate. Third, an exact historical match is reused when available; otherwise, open-space retrieval transfers bounded residual evidence from protocol-visible historical states through explicit relations over difficulty, temporal suffixes, ancestors, and outcome structure. Each transferred correction retains its source, relation, and weight, yielding a traceable path from historical evidence to the final half-life prediction. Accurate prediction alone does not decide which items should be reviewed when many are due under a fixed daily budget. The scheduler must react when outcomes show overconfidence, rank competing due items by the harm of additional delay, and prevent long-tail starvation. We therefore introduce the OSR Adaptive Calibration-Regret Index (OSR-ACRI), which converts OSR-HLR estimates into due dates and closes the loop between memory prediction and capacity-aware scheduling.

Our contributions are fourfold: ❶ **Open-space retrieval for half-life prediction.** We introduce OSR-HLR, which combines relative half-life estimation with deterministic residual transfer over recurring and related memory states. Historical evidence comes from protocol-visible sources, and every correction remains traceable to its source state and relation. ❷ **Capacity-aware scheduling.** OSR-ACRI allocates a fixed daily budget by combining one-sided calibration, the marginal harm of delay, and queue age. ❸ **Prediction evaluation.** Under source-only protocols that keep every Test outcome out of training and retrieval, OSR-HLR achieves the lowest H-SMAPE among 27 baselines in all 15 dataset–protocol settings. Its Row-protocol reductions are **17.40%/2.59%/42.46%** on MaiMemo/Anki/EdNet-Recall, and it leads on held-out Anki users. ❹ **Decision evaluation.** With the scheduler fixed, OSR-HLR reduces 365-day regret by **8.7%–15.8%**. With the predictor and workload fixed, OSR-ACRI improves retention AUC by **11.1%** and final recall by **19.8%** and reduces time below target by **7.5%** relative to Threshold.

## 2 Related Work

Spaced-repetition prediction spans compact cognitive or stability-update models (Settles & Meeder, 2016; Lindsey et al., 2014; Anderson et al., 1997; Ye et al., 2022; Su et al., 2023), neural forgetting and knowledge-tracing models that encode ordered learner histories (Cho et al., 2014; Piech et al., 2015; Nagatani et al., 2019; Ma et al., 2023; Xiao & Wang, 2024; Shen et al., 2026a; Sun et al., 2025; Guan et al., 2025; Shao et al., 2025), and graph-based models that represent typed dependencies or transfer among learning entities (Schlichtkrull et al., 2018; Hu et al., 2020; Cui et al., 2024; Xu et al., 2026). HLR estimates log half-life from practice statistics, while DASH incorporates difficulty and spacing (Settles & Meeder, 2016; Lindsey et al., 2014). Forgetting-aware models further incorporate elapsed-time features or attention (Ma et al., 2023; Guan et al., 2025). OSR-HLR differs by combining relative half-life prediction with explicit state-level evidence reuse. It aggregates recurring memory states and transfers residual evidence through deterministic relations whose historical sources remain visible throughout inference. On the policy side, Threshold, DHP, SSP-MMC, MEMORIZE (Tabibian et al., 2019), and Anki-style schedulers encode different review objectives (Ye et al., 2022; Su et al., 2023), but they usually do not jointly handle online predictor calibration and competition among many due items under a hard daily budget. OSR-ACRI addresses that capacity-constrained setting while keeping the predictor fixed, so the scheduling effect can be measured separately from prediction accuracy. Appendix A gives the expanded comparison.

# 3 PROBLEM FORMULATION

**Prediction task.** After $L$ reviews, a learner–item history contains an outcome sequence $\mathbf{r} = (r_1, \ldots, r_L)$ and an interval sequence $\mathbf{t} = (t_1, \ldots, t_L)$, where $t_k$ is the delay since the preceding review. Following prior half-life and stability models that condition memory dynamics on difficulty (Ye et al., 2022; Su et al., 2023; Lin & Hu, 2024), the task uses item difficulty $d$, query delay $\Delta t$, and count-based support features $\mathbf{c}$ to predict

$$\hat{h} = f_\theta(d, \mathbf{r}, \mathbf{t}, \Delta t, \mathbf{c}), \tag{1}$$

**Aggregated observations and half-life labels.** Following prior half-life construction (Settles & Meeder, 2016; Ye et al., 2022; Su et al., 2023), reviews with the same history and query interval form one aggregated observation $(d, \mathbf{r}, \mathbf{t}, \Delta t_i, c_i, R_i)$, where $R_i = a_i/c_i$ is the observed recall rate from $c_i$ reviews. The memory state is $v = (d, \mathbf{r}, \mathbf{t})$; excluding $\Delta t_i$ lets several query delays identify points on the same forgetting curve. We fit one state-level label by count-weighted regression:

$$\log R_i = -\frac{\log 2}{h_v}\Delta t_i, \qquad k_v = \frac{\sum_i c_i \Delta t_i \log R_i}{\sum_i c_i \Delta t_i^2}, \qquad h_v = -\frac{\log 2}{k_v}. \tag{2}$$

Larger groups receive more weight because their recall rates are estimated from more reviews. Section 5.1 gives the dataset-specific outcome coding, information boundary, and Train-only regularization details.

# 4 METHOD

## 4.1 OVERVIEW

As shown in Figure 2, the method links a half-life predictor with a capacity-aware scheduler. Open-Space Retrieval for Half-Life Regression (OSR-HLR) treats half-life prediction as relative estimation followed by state-level evidence transfer. It first estimates memory strength from the current history and query, then aggregates recurring states. It reuses exact historical matches when available and otherwise retrieves residual evidence from related protocol-visible states. OSR Adaptive Calibration-Regret Index (OSR-ACRI) converts these estimates into review decisions by updating due dates after overconfident outcomes and ranking due items by one-day delay harm plus queue age.

## 4.2 OPEN-SPACE RETRIEVAL FOR HALF-LIFE REGRESSION (OSR-HLR)

**Model inputs and prediction path.** OSR-HLR maps each aggregated observation to a half-life prediction from four inputs: the difficulty bin $d_i$, a recent-event matrix $\mathbf{X}_i \in \mathbb{R}^{T_{\max} \times F_e}$, its padding mask $\mathbf{m}_i$, and a static query vector $\mathbf{s}_i \in \mathbb{R}^{F_s}$. Here $T_{\max}$ is the maximum number of retained past reviews, $F_e$ is the number of features per review event, and $F_s$ is the number of static query features. The event features encode outcome direction, spacing, position, and available historical memory strength, while the static vector summarizes query support and multiscale transition behavior.

The observation encoder predicts memory strength relative to the query interval:

$$o_i = g_\theta(\mathbf{X}_i, \mathbf{m}_i, \mathbf{s}_i, d_i), \qquad q_i = o_i + \log \Delta t_i = \log \hat{h}_i. \tag{3}$$

Observations with the same state key $v = (d, \mathbf{r}, \mathbf{t})$ are then aggregated, after which a correction layer may use historical evidence permitted by the evaluation protocol:

$$(\mathbf{X}_i, \mathbf{m}_i, \mathbf{s}_i, d_i) \longrightarrow q_i \longrightarrow \hat{h}_v^{(0)} \longrightarrow \hat{h}_v^{(\text{final})}. \tag{4}$$

This separates sequence learning, reconciliation of repeated observations, and auditable evidence transfer. Figure 2 summarizes the pipeline.

**Relative half-life base estimator.** Projected event features are encoded by a masked bidirectional GRU. Attention combines a learned content score with an outcome-aware basis for recency, outcome transitions, streaks, relative spacing, and last-forget/recovery structure, with depth gates suppressing terms unsupported by short histories. The attention summary, terminal GRU states, projected static vector, and difficulty embedding are fused and passed through a main head plus bounded local correction heads. Predicting $o_i = \log(h_i/\Delta t_i)$ converts multiplicative half-life variation into an additive offset, with $\hat{h}_i = \exp(o_i + \log \Delta t_i)$. The feature and encoder details are in Appendix E.

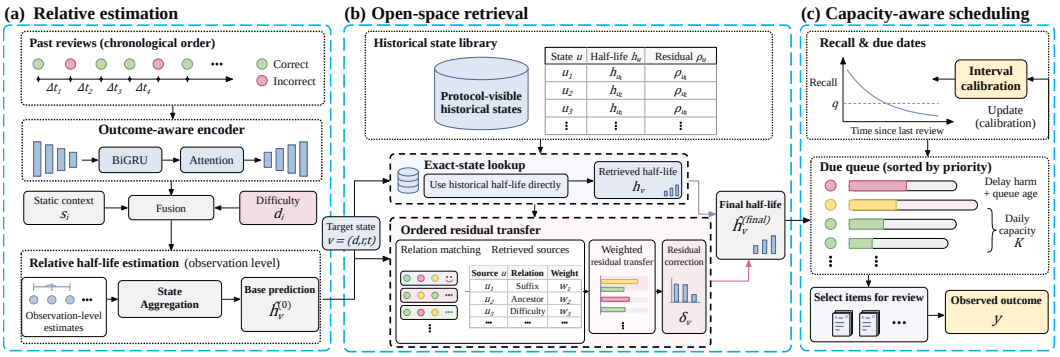


Figure 2: Overview of the framework. (a) OSR-HLR estimates half-life relative to the query interval and aggregates it by memory state. (b) Open-space retrieval reuses exact historical matches or transfers residual evidence from related states. (c) OSR-ACRI converts these estimates into review decisions under a fixed daily capacity.

**Open-space residual retrieval.** For observation predictions $q_j = \log \hat{h}_j$ sharing state key $v$, the state-level base is

$$b_v = \frac{\sum_{j \in \mathcal{I}_v} \sqrt{c_j}\, q_j}{\sum_{j \in \mathcal{I}_v} \sqrt{c_j}}, \qquad \hat{h}_v^{(0)} = \exp(b_v). \tag{5}$$

Square-root weighting gives better-supported observations more influence without allowing one aggregate to dominate linearly; within-state dispersion is retained as source uncertainty.

For evaluation split $S$, let $\mathcal{V}_S$ contain only historical states whose fitted half-lives are available under the protocol. A state matching one of these records reuses its historical half-life; otherwise it retains the base prediction:

$$\hat{h}_v^{(\text{lookup})} = \begin{cases} h_v, & v \in \mathcal{V}_S, \\ \hat{h}_v^{(0)}, & v \notin \mathcal{V}_S. \end{cases} \tag{6}$$

The held-out outcome currently being predicted is never included among these historical sources. For targets without an exact historical match, an ordered family of relations $\mathcal{R} = (r_1, \ldots, r_K)$ aggregates visible source residuals:

$$\begin{aligned} \rho_u &= \log h_u - \log \hat{h}_u^{(0)}, \\ m_v^{(k)} &= \frac{\sum_{u \in \mathcal{N}_k(v)} w_{uv}^{(k)}\, \rho_u}{\sum_{u \in \mathcal{N}_k(v)} w_{uv}^{(k)}}, \qquad \hat{h}_v^{(k)} = \hat{h}_v^{(k-1)} \exp\Big(\alpha_k \operatorname{clip}(m_v^{(k)})\Big). \end{aligned} \tag{7}$$

The relation family covers exact-history transfer across difficulty, short and long temporal suffixes, ancestors, difficulty trends, historical parent measurements, and outcome-sign fallbacks. Weights combine source support, relation distance, source disagreement, and uncertainty. Source residuals remain anchored to the unchanged neural base while bounded corrections accumulate on the target.

This layer is a deterministic residual chain rather than a learned GNN. Its relation order, source mask, support indicator, weight rule, clipping bound, and coefficient are fixed and inspectable. Consequently every correction can be traced to the earlier observations permitted by the evaluation protocol, which is the sense in which the final layer is auditable.

**Training objective.** The neural predictor is trained before deterministic state correction and scheduling. A single objective combines relative half-life error, robust log-space error, and optional recall consistency:

$$\begin{aligned} \mathcal{L}(\theta) = &\frac{\sum_{i \in \mathcal{B}} w_i \Big[\lambda_h \frac{2|\hat{h}_i - h_i|}{\hat{h}_i + h_i + \varepsilon} + (1 - \lambda_h)\, \mathrm{SmoothL1}_\beta(\log \hat{h}_i, \log h_i)\Big]}{\sum_{i \in \mathcal{B}} w_i} \\ &+ \lambda_R \frac{\sum_{i \in \mathcal{B}} \tilde{w}_i\, \mathrm{BCE}(R_i, 2^{-\Delta t_i / \hat{h}_i})}{\sum_{i \in \mathcal{B}} \tilde{w}_i}. \end{aligned} \tag{8}$$

Here $w_i$ and $\tilde{w}_i$ are half-life and recall reliability weights, respectively, and $\lambda_R$ controls the optional Train-only recall consistency term. $R_i$ is never passed to the forward function or available during Validation or Test inference.

### 4.3 OSR Adaptive Calibration-Regret Index (OSR-ACRI)

OSR-ACRI converts OSR-HLR predictions into review decisions under a hard daily capacity. It begins with the target-recall interval $-\hat{h}_{i,t} \log_2 q$, where the target recall $q$ is the intended post-review recall probability, and maintains an exponential moving average $c_t$ of observed calibration residuals. Only positive residuals, which indicate overconfidence, contract future intervals:

$$c_t = \gamma c_{t-1} + (1-\gamma)|\mathcal{B}_t|^{-1} \sum_{j \in \mathcal{B}_t} (\hat{p}_j - y_j),$$
$$g_t = \exp(-\kappa [c_t]_+), \qquad \Delta_{i,t} = \max\{1, \mathrm{round}(g_t[-\hat{h}_{i,t} \log_2 q])\}. \tag{9}$$

When more items are due than capacity permits, OSR-ACRI ranks item $i$ by

$$I_i(t) = (t - d_i) + \frac{\beta g_t}{1-q} \left( [q - \hat{p}_i(a_i + 1)]_+ - [q - \hat{p}_i(a_i)]_+ \right). \tag{10}$$

The first term prevents starvation; the second is the predicted increase in target shortfall from waiting one more day. After each review, the observed outcome updates the history, OSR-HLR re-predicts half-life, and OSR-ACRI updates calibration and the due queue. This closes the simulated feedback loop. The complete daily loop is listed in Appendix G.

## 5 Experiments

### 5.1 Experimental Setup

**Datasets.** We evaluate on three datasets. MaiMemo (Su et al., 2023) records binary review outcomes; the public Anki review-log release contains 10,000 user collections and uses rating 1 for forgetting and ratings 2–4 for recall; and EdNet-Recall, which we derive from EdNet-KT1 (Choi et al., 2020), uses delayed repeated question responses from a public adaptive TOEIC practice log. Table 1 summarizes source scale and the source-only Row-protocol data used for the headline three-dataset comparison.

**Baselines and inputs.** We compare OSR-HLR with 27 baselines spanning formulaic SRS rules, generic sequence models, memory/forgetting models, and recent knowledge-tracing models. All methods are evaluated under the same input-information boundary rather than identical feature vectors. Each method may construct its model-specific representation from past outcomes, past intervals, item difficulty, query delay, and support counts. Only in the transductive label-pool view does OSR-HLR additionally read the released per-event strength history and the state multiplicity; all source-only protocols disable both. No method receives the query response or aggregated recall rate as an inference input; these quantities are used only as supervised targets and evaluation labels. Complete model details are in Appendix F.3.

**Prediction setting.** Five source-only generalization protocols (Row/Learner/Item/State/Joint) share a common Train/Validation source boundary; the Row protocol gives the headline three-dataset comparison. In every source-only protocol, half-life labels are refitted within each split, so no Train or Validation label is fitted from Test reviews and exact lookup cannot reproduce a Test target. Held-out Anki users, also with split-local labels, provide a separate identity and chronological confirmation. MaiMemo and Anki, the two native spaced-repetition system (SRS) logs, use ten retrainings, and EdNet-Recall uses three. EdNet-Recall uses split-local weighted exponential compatibility projections because EdNet is not a native SRS log. Its item difficulty is estimated only from Train events using a learner-disjoint nuisance fold; held-out items use metadata fallback rather than held-out outcomes. Split assignments do not depend on outcomes. In the source-only protocols, each held-out outcome is used only for scoring and does not enter training, retrieval, residual correction, or model selection.

**Prediction metrics.** The main prediction table reports H-SMAPE for relative half-life error, H-MAE for day-scale half-life error, and P-RMSE for recall-probability error. Full feature, training, protocol, and metric details are in Appendices E, F, G, and H.

Table 1: Dataset scale and primary source-only Row-protocol modeling data. Source records are pre-aggregation reviews/interactions; Anki items are learner–card pairs; states are counted within each split. The transductive label-pool view in the appendix uses the full released MaiMemo and Anki tables (71,696 and 106,726 observations).

| Dataset | Source records | Users | Items/Qs | Model rows | States | Train | Val | Test |
|---|---|---|---|---|---|---|---|---|
| MaiMemo | 226.4M | 660k | 17.1k | 12,983 | 12,696 | 7,348 | 2,802 | 2,833 |
| Anki | 727.7M | 10k | 95.6M | 12,231 | 5,565 | 8,511 | 2,089 | 1,631 |
| EdNet-Recall | 95.3M | 784k | 12.2k | 20,322 | 1,703 | 9,921 | 5,235 | 5,166 |

**Prediction-to-scheduling outcomes (Section 5.3).** The setting fixes one shared scheduler and varies only the predictor. The comparators are the 27 prediction baselines and OSR-HLR, all evaluated under the same simulated states, transition environments, capacity, and random streams. The metrics are next-review target error, recall shortfall, early-review surplus, and four 365-day regret measures relative to a perfect-prediction reference. This isolates whether better half-life prediction improves decisions without changing the scheduling rule. Complete simulator details are in Appendix G.

**Capacity-aware scheduling (Section 5.4).** The setting fixes a three-seed OSR-HLR ensemble and varies only the scheduler. The baselines are six reference schedulers evaluated under matched states, source-domain transition environments, daily capacity, and random streams. The metrics are retention AUC, time below the target recall, final recall, and review workload. This isolates whether the capacity-aware index improves scheduling once prediction is fixed. Section 4.3 gives the algorithm; complete Validation-selected settings and component controls are in Appendix G.

### 5.2 Prediction Accuracy and Generalization

**Prediction accuracy.** Table 2 reports our primary comparison, the source-only Row protocol. Labels are refitted within each split, residual evidence comes only from Train and Validation states, historical-strength and state-multiplicity inputs are disabled, and every baseline receives the same residual postprocessor as OSR-HLR. Under this protocol, OSR-HLR reduces H-SMAPE by **17.40%**, **2.59%**, and **42.46%** on MaiMemo, Anki, and EdNet-Recall, respectively, relative to the strongest baseline in each column. Open-space retrieval is model-agnostic: it lowers H-SMAPE for all 27 baselines on all three datasets, with median relative reductions of 8.3%, 11.6%, and 4.9% (e.g., from 17.44 to 16.26 for LTKT on MaiMemo; Appendix I). Because every baseline receives this retrieval layer, the remaining gap reflects the relative-target base estimator of OSR-HLR. Appendix H also reports a transductive label-pool view on the full released tables. We do not use it for headline claims because released labels are shared by all observations of a state, so exact lookup can reproduce a Test target. OSR-HLR is best in all metric cells across all datasets, against different strongest competitors. Its margins are largest on MaiMemo and EdNet-Recall, where it also reduces day-scale H-MAE by **29.41%** and **47.69%**. The smaller Anki and EdNet P-RMSE gains are compatible with the fact that $R(\Delta t) = 2^{-\Delta t/h}$ compresses half-life differences in high-recall regions. The following core ablations examine the two mechanisms underlying OSR-HLR: relative half-life estimation and open-space residual transfer.

**Five source-only generalization protocols.** Figure 3 compares the five source-only generalization protocols (Row/Learner/Item/State/Joint) using the same correction procedure, restricted to evidence from Train and Validation. Across all 27 baselines, OSR-HLR has the lowest mean H-SMAPE in all fifteen dataset–protocol settings. Reductions are **14.35–18.55%** on MaiMemo, **1.54–6.40%** on Anki, and **14.05–42.46%** on EdNet-Recall. The smallest native-SRS effects are the Anki row, item, and joint settings. MaiMemo remains in a narrow positive range as identity overlap is removed, whereas Anki gains are smaller and vary more. Because labels are refitted within each split, no Train or Validation label contains Test reviews, so the gains cannot come from copying Test labels. Joint is the strictest source-domain setting because it removes learner, item, and complete-state overlap; it is not cross-dataset transfer. Full results and protocol details are in Appendix D and Table 7.

**Independent confirmation on held-out Anki users.** We used 10% of Anki users for development and reserved the remaining 90% for confirmation. Against the prespecified six-model family, OSR-HLR lowers H-SMAPE by **13.37%** under identity isolation and **4.20%** under within-learner

Table 2: Prediction performance on the three benchmarks. Values are mean ± sample standard deviation; lower is better. **Red bold**, blue underlined, and orange values mark the best, second-best, and third-best results in each column. Improvement is relative to the best baseline in each column.

| Method | MaiMemo | | | Anki | | | EdNet-Recall | | |
|---|---|---|---|---|---|---|---|---|---|
| | H-SMAPE ↓ | H-MAE ↓ | P-RMSE ↓ | H-SMAPE ↓ | H-MAE ↓ | P-RMSE ↓ | H-SMAPE ↓ | H-MAE ↓ | P-RMSE ↓ |
| *Cognitive / formulaic* | | | | | | | | | |
| Pimsleur | 63.79±0.07 | 50.34±0.06 | 0.2255±0.0003 | 100.02±0.01 | 178.77±0.08 | 0.3051±0.0000 | 135.46±0.01 | 300.59±0.02 | 0.3236±0.0000 |
| Leitner | 61.15±0.01 | 36.34±0.01 | 0.1418±0.0000 | 94.93±0.00 | 180.78±0.02 | 0.3084±0.0000 | 137.52±0.00 | 308.86±0.01 | 0.3291±0.0000 |
| ACT-R | 51.70±0.04 | 41.05±0.29 | 0.0969±0.0002 | 73.73±0.05 | 137.62±0.24 | 0.2400±0.0001 | 77.52±0.02 | 251.95±0.07 | 0.2906±0.0000 |
| HLR | 50.29±1.86 | 39.37±2.38 | 0.0884±0.0061 | 70.44±2.03 | 123.16±3.67 | 0.2432±0.0050 | 74.04±0.81 | 221.28±3.91 | 0.2844±0.0005 |
| DASH | 42.29±1.12 | 35.94±1.31 | 0.0954±0.0038 | 70.05±1.44 | 122.63±3.14 | 0.2417±0.0030 | 75.67±0.69 | 219.62±1.84 | 0.2891±0.0036 |
| DHP | 47.65±0.02 | 32.26±0.03 | 0.1056±0.0000 | 72.06±0.07 | 121.29±0.11 | 0.2318±0.0001 | 82.94±3.55 | 269.34±23.62 | 0.2951±0.0067 |
| FSRS | 32.63±0.40 | 18.28±0.30 | 0.0699±0.0030 | 52.25±1.33 | 98.67±3.75 | 0.1798±0.0024 | 69.83±6.18 | 208.00±16.02 | 0.2783±0.0086 |
| *Generic sequence* | | | | | | | | | |
| RNN | 17.56±0.73 | 14.45±1.07 | 0.0351±0.0012 | 40.77±1.04 | 82.76±5.83 | 0.1739±0.0008 | 60.46±0.28 | 194.28±6.55 | 0.2718±0.0009 |
| GRU | 16.81±0.21 | 13.22±0.20 | 0.0336±0.0005 | 41.21±0.78 | 83.20±4.80 | 0.1748±0.0004 | 72.05±1.27 | 213.68±5.71 | 0.2876±0.0037 |
| LSTM | 17.03±0.48 | 13.37±0.41 | 0.0345±0.0015 | 40.91±0.79 | 80.89±4.37 | 0.1747±0.0005 | 69.30±8.10 | 208.40±25.49 | 0.2848±0.0101 |
| Transformer | 17.55±0.70 | 14.06±1.09 | 0.0344±0.0013 | 42.17±1.56 | 91.42±4.66 | 0.1714±0.0035 | 70.59±3.79 | 228.88±19.35 | 0.2807±0.0023 |
| DKT | 16.61±0.44 | 13.06±0.44 | 0.0338±0.0010 | 40.12±0.85 | 83.54±2.11 | 0.1733±0.0023 | 61.40±0.63 | 186.05±1.38 | 0.2741±0.0007 |
| *Memory / forgetting* | | | | | | | | | |
| FIFAKT | 16.80±0.47 | 13.49±0.44 | 0.0332±0.0010 | 39.34±1.01 | 82.24±3.08 | 0.1732±0.0014 | 62.17±0.62 | 201.38±7.99 | 0.2728±0.0015 |
| GRU-HLR | 16.68±0.17 | 13.05±0.15 | 0.0335±0.0005 | 40.88±0.82 | 80.19±2.78 | 0.1731±0.0008 | 59.26±0.83 | 183.29±5.65 | 0.2721±0.0005 |
| THLR | 17.51±0.32 | 13.37±0.21 | 0.0349±0.0009 | 42.86±1.68 | 86.20±1.96 | 0.1706±0.0048 | 62.72±0.54 | 190.13±7.29 | 0.2756±0.0012 |
| ROKAN | 31.38±5.05 | 30.16±5.90 | 0.1401±0.0793 | 47.48±1.76 | 86.31±3.67 | 0.1820±0.0019 | 64.31±0.44 | 206.97±4.57 | 0.2740±0.0008 |
| I-Mem | 18.94±1.00 | 15.83±1.60 | 0.0378±0.0017 | 40.96±0.90 | 85.22±2.09 | 0.1743±0.0009 | 60.73±0.71 | 187.43±3.42 | 0.2730±0.0012 |
| *Recent KT / memory* | | | | | | | | | |
| EPsyINN | 16.28±0.22 | 12.97±0.21 | 0.0323±0.0003 | 40.33±0.69 | 83.87±3.50 | 0.1722±0.0014 | 61.24±0.14 | 184.63±1.58 | 0.2739±0.0002 |
| LefoKT | 17.66±0.36 | 13.90±0.34 | 0.0370±0.0012 | 42.34±2.09 | 90.38±4.41 | 0.1680±0.0019 | 66.91±0.51 | 199.53±2.42 | 0.2781±0.0007 |
| ReKT | 17.09±0.24 | 13.48±0.23 | 0.0340±0.0005 | 41.53±0.73 | 86.34±1.92 | 0.1707±0.0030 | 61.64±0.37 | 187.91±3.00 | 0.2742±0.0007 |
| LTKT | 16.26±0.51 | 13.19±0.53 | 0.0328±0.0013 | 40.74±1.23 | 81.19±5.06 | 0.1725±0.0014 | 62.18±0.74 | 186.80±1.14 | 0.2746±0.0014 |
| SPsyINN | 16.68±0.42 | 13.22±0.34 | 0.0330±0.0008 | 39.74±1.03 | 80.59±5.80 | 0.1737±0.0013 | 59.99±0.74 | 182.81±2.70 | 0.2720±0.0009 |
| KVFKT | 17.35±0.43 | 13.46±0.23 | 0.0349±0.0010 | 43.63±0.71 | 88.82±1.66 | 0.1736±0.0006 | 66.69±7.19 | 204.89±27.87 | 0.2795±0.0070 |
| GRKT | 17.71±0.39 | 14.19±0.61 | 0.0358±0.0010 | 42.07±1.50 | 86.69±2.67 | 0.1717±0.0008 | 71.96±7.71 | 220.47±26.81 | 0.2893±0.0117 |
| MIKT | 16.68±0.38 | 13.15±0.37 | 0.0332±0.0006 | 41.70±0.80 | 84.15±4.81 | 0.1710±0.0010 | 62.32±0.79 | 187.45±1.86 | 0.2751±0.0010 |
| MCKT | 16.47±0.27 | 13.13±0.38 | 0.0329±0.0005 | 40.35±1.10 | 83.51±3.51 | 0.1719±0.0008 | 60.83±0.87 | 186.88±4.20 | 0.2729±0.0015 |
| DGMKT | 16.83±0.23 | 13.22±0.17 | 0.0335±0.0005 | 39.38±1.35 | **76.63±6.87** | 0.1714±0.0027 | 60.99±0.93 | 186.20±2.32 | 0.2732±0.0019 |
| *Ours* | | | | | | | | | |
| **OSR-HLR** | **13.43**±0.05 | **9.15**±0.05 | **0.0272**±0.0001 | **38.32**±1.00 | **73.69**±2.60 | **0.1611**±0.0028 | **34.10**±0.08 | **95.62**±3.61 | **0.2511**±0.0007 |
| **Improvement (%)** | **17.40** | **29.41** | **15.93** | **2.59** | **3.84** | **4.06** | **42.46** | **47.69** | **7.61** |

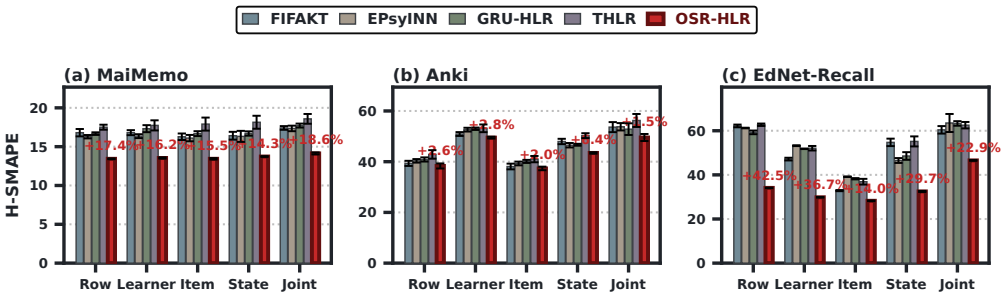


Figure 3: Source-only protocols: H-SMAPE with representative baselines.

chronological evaluation (Figure 4). The 27-baseline extension is also positive against the strongest per-protocol baselines: **11.36%** under identity isolation (ReKT) and **2.40%** chronologically (LefoKT). Full results are in Appendix C. Positive results under both identity isolation and chronological evaluation rule out gains confined to repeated users or random interleaving.

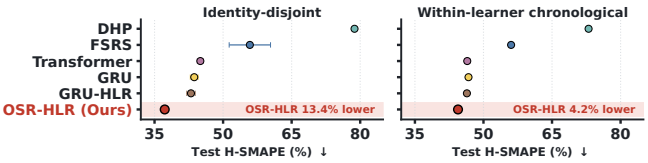


Figure 4: Held-out Anki users: mean Test H-SMAPE over ten seeds; bars show sample SD.

Table 3: Next-review reduction (%).

| $q$ | Target error | Recall shortfall | Early-rev. surplus |
|---|---|---|---|
| 0.70 | **19.5** | **22.6** | **15.9** |
| **0.90** | **19.7** | **23.1** | **15.6** |
| 0.95 | **19.8** | **23.2** | **15.5** |

**Core-mechanism ablation.** We isolate the two components most directly tied to the proposed prediction mechanism. In Figure 5a, *w/o Relative Target*, *w/o Residual Graph*, and *w/o Both Core Mechanisms* are first shown as absolute Test H-SMAPE increases and then normalized in Figure 5b as relative increases. The absolute changes are **0.352/0.269/1.007** points on MaiMemo, **2.217/3.353/6.398** points on Anki, and **0.874/0.919/2.109** points on EdNet-Recall, respectively. All

cells use source-only protocols: MaiMemo and Anki use the Row protocol with ten seeds, and EdNet-Recall averages the five source-only protocols. All paired 95% intervals exclude zero across the nine dataset-control comparisons. Removing both mechanisms hurts more than either removal alone: relative targeting normalizes memory scale, while residual transfer corrects errors shared by related states. Their interaction is positive on MaiMemo (**+0.386** points, 95% CI $[+0.312, +0.460]$) and Anki (**+0.829**, $[+0.477, +1.180]$), so the two mechanisms are complementary rather than independently additive. Two illustrative decision corrections are shown in Appendix Figure 9; full ablation, attribution, and case-selection details are in Appendix B.

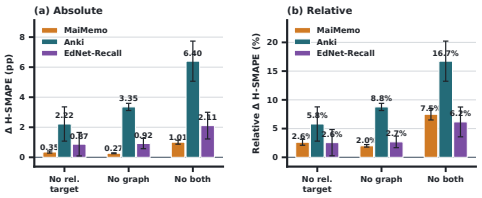


Figure 5: Core ablation scales.

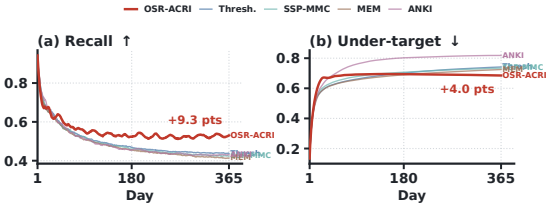


Figure 6: Daily scheduling trajectory under matched capacity.

### 5.3 From Prediction to Scheduling Outcomes

Holding the scheduler fixed, we test whether OSR-HLR's prediction gains transfer into better review decisions and lower long-horizon regret.

**Next-review decisions.** Every predictor feeds the same threshold scheduler, and each chosen interval is scored against an independently fitted reference half-life. At $q = 0.90$, OSR-HLR reduces target-recall error, late-review shortfall, and early-review recall surplus by **19.73%**, **23.12%**, and **15.60%**, respectively, against the matched EPsyINN predictor with the same postprocessor. Against the strongest of all 27 baselines for each metric, the reductions are **19.5%**, **20.8%**, and **15.6%**; the direction persists across all six targets (Table 10). Shortfall improves most, so the effect is stronger for late than unnecessary early reviews. The shared rule attributes this difference to interval estimates, not scheduler choice. Metric definitions, bootstrap inference, and the complete sensitivity ranges are in Appendix G.

**Prediction error and long-horizon regret.** To isolate prediction quality, every method uses the same $q = 0.90$ integer-day scheduler. Measured held-out errors are propagated through matched source-fitted simulations and compared with a perfect prediction reference using the same rule. This reference is neither learned nor globally optimal. Full simulator, error-transport, and aggregation details are in Appendix G.

Figure 7 shows four cumulative regrets over 365 days. OSR-HLR is lower at every reported checkpoint after day 0 through the full year against the corresponding strongest baseline.

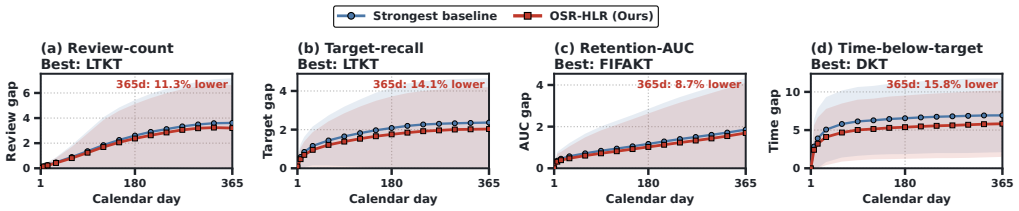


Figure 7: Cumulative regret relative to a perfect-prediction reference under the shared threshold scheduler. Each panel compares OSR-HLR with the strongest baseline for that metric after aggregation over four scenarios; lower is better.

Across all 27 baselines and four prespecified scenarios, covering 108 baseline–scenario comparisons, OSR-HLR reduces review-count, target-recall, retention-AUC, and time-below-target regret relative to the strongest baseline by **11.3%**, **14.1%**, **8.7%**, and **15.8%**, respectively (Table 11). Held-out prediction error also closely tracks long-horizon regret across all 27 baselines ($\rho = 0.98$; Figure 12). This links the one-step and year-long results: repeated interval errors shift review dates and accumulate

into regret. The fixed scheduler rules out policy changes, but the result remains simulator-based and does not establish synthetic-history robustness, global optimality, or causal learner benefit.

### 5.4 OSR-ACRI Improves Scheduling under a Fixed Predictor

We fix a three-seed OSR-HLR ensemble and vary only the scheduler under matched states, environments, capacity, and random streams. OSR-ACRI is the scheduler defined by the one-sided online calibration, marginal-delay index, and queue-age terms in Section 4.3; its settings are selected on disjoint Validation states before Test outcomes are opened. Full evaluation details are in Appendix G. Relative to the workload-matched Threshold scheduler, OSR-ACRI uses the same 11,680 reviews while improving retention AUC by **11.1%** and final recall by **19.8%** and reducing time below target by **7.5%** (Table 24). Among the workload-matched references, MEMORIZE is stronger on time below target, against which the reduction is **6.2%**; all aggregate metric comparisons still favor OSR-ACRI, with consistent source-domain environment behavior. Matched workload rules out simply allocating more reviews: OSR-ACRI instead changes which due items receive capacity under contention. Under the identical daily capacity and total review budget, Figure 6 shows OSR-ACRI holding recall roughly nine points higher across the second half of the year while spending the least time below target; the day-365 means reproduce the frozen Test endpoints. This supports a source-domain, capacity-constrained scheduling result, not global optimality, cross-source portability, or causal learner benefit. Figure 8 shows that dropping the delay index or queue age degrades all three utility metrics. Removing the delay index nearly reproduces Threshold, identifying marginal delay cost as the main ranking signal; removing queue age causes the largest aggregate loss, establishing that age information matters without directly measuring starvation. Removing calibration gives nearly the same AUC and time below target and slightly higher final recall, so calibration is a minor contributor to these endpoints rather than an established safety mechanism.

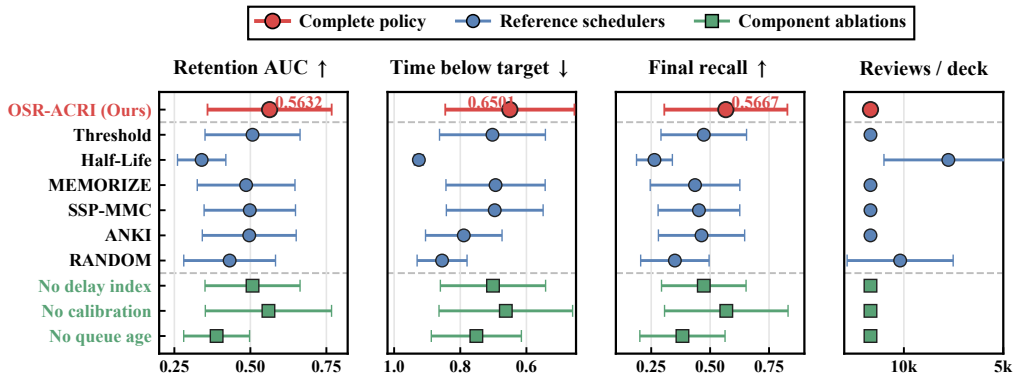


Figure 8: Fixed-predictor comparison of OSR-ACRI with reference schedulers and component ablations. Points show test means and horizontal bars indicate sample standard deviations. Arrows denote the preferred direction for each metric. Details are shown in Table 24.

## 6 Conclusion

We presented a closed-loop framework that couples half-life prediction with capacity-aware review scheduling. OSR-HLR predicts half-life relative to the query interval, which separates memory strength from query timing, and aggregates recurring observations into memory states. Its open-space retrieval then transfers residual evidence from protocol-visible historical states through ordered, deterministic relations, so every correction can be traced to its source state, relation, and weight. OSR-ACRI converts these estimates into review decisions under a hard daily budget by combining one-sided calibration, the marginal harm of delay, and queue age. Across three benchmarks, OSR-HLR achieves the lowest H-SMAPE among 27 baselines in all 15 source-only settings, open-space retrieval improves every baseline as a model-agnostic component, and OSR-ACRI raises retention and final recall at matched workload. A prospective study with real learners is the natural next step toward validating this loop in deployed spaced-repetition systems.

## AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experiment design; assist with implementation and data-processing code; support result analysis and interpretation; and assist with manuscript drafting, editing, figure and table preparation, literature search, and reference formatting. We did not use generative AI to generate synthetic participant data, formulate mathematical proofs, or replace the reported empirical evaluation. The authors inspected the generated code, executed the reported tests and audits, checked numerical claims against frozen machine-readable artifacts, and verified cited works against publisher, proceedings, DOI, or archival records. We reviewed all AI-assisted work and take responsibility for the final content, claims, and artifacts.

## REFERENCES

John R Anderson, Michael Matessa, and Christian Lebiere. Act-r: A theory of higher level cognition and its relation to visual attention. *Human-computer interaction*, 12(4):439–462, 1997. doi: 10.1207/s15327051hci1204_5.

Youheng Bai, Xueyi Li, Zitao Liu, Yaying Huang, Mi Tian, and Weiqi Luo. Rethinking and improving student learning and forgetting processes for attention based knowledge tracing models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 27822–27830, 2025. doi: 10.1609/aaai.v39i27.34998.

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 the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179/.

Youngduck Choi, Youngnam Lee, Dongmin Shin, Junghyun Cho, Seoyon Park, Seewoo Lee, Jineon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. Ednet: A large-scale hierarchical dataset in education. In *International conference on artificial intelligence in education*, pp. 69–73. Springer, 2020.

Jiajun Cui, Hong Qian, Bo Jiang, and Wei Zhang. Leveraging pedagogical theories to understand student learning process with graph-based reasonable knowledge tracing. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 502–513, 2024. doi: 10.1145/3637528.3671853.

Hermann Ebbinghaus. *Memory: A Contribution to Experimental Psychology*. Teachers College, Columbia University, New York, NY, 1913. URL https://archive.org/details/memorycontributi00ebbiuoft.

Jeffrey L Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990. doi: 10.1207/s15516709cog1402_1.

Quanlong Guan, Xiuliang Duan, Kaiquan Bian, Guanliang Chen, Jianbo Huang, Zhiguo Gong, and Liangda Fang. KVFKT: A new horizon in knowledge tracing with attention-based embedding and forgetting curve integration. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4399–4409, 2025. doi: 10.18653/v1/2025.coling-main.297. URL https://aclanthology.org/2025.coling-main.297/.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*, pp. 2704–2710, 2020. doi: 10.1145/3366423.3380027. URL https://arxiv.org/abs/2003.01332.

Sebastian Leitner. *So lernt man lernen:[angewandte Lernpsychologie-ein Weg zum Erfolg]*. Herder, 1972.

Gaoyun Lin and Yahong Hu. Exploring cognitive diversity and dynamics for effective language memory retention. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024. URL https://escholarship.org/uc/item/2q10r2x5.

Robert V Lindsey, Jeffery D Shroyer, Harold Pashler, and Michael C Mozer. Improving students' long-term knowledge retention through personalized review. *Psychological science*, 25(3): 639–647, 2014. doi: 10.1177/0956797613504302.

Boxuan Ma, Gayan Prasad Hettiarachchi, Sora Fukui, and Yuji Ando. Each encounter counts: Modeling language learning and forgetting. In *Proceedings of the 13th International Learning Analytics and Knowledge Conference*, pp. 79–88, 2023. doi: 10.1145/3576050.3576062.

Boxuan Ma, Sora Fukui, Yuji Ando, and Shin'ichi Konomi. Integrating forgetting behavior and linguistic features in language learning models. *ACM Transactions on Knowledge Discovery from Data*, 20(2):1–26, 2026. doi: 10.1145/3778163.

Koki Nagatani, Qian Zhang, Masahiro Sato, Yan-Ying Chen, Francine Chen, and Tomoko Ohkuma. Augmenting knowledge tracing by considering forgetting behavior. In *The world wide web conference*, pp. 3101–3107, 2019. doi: 10.1145/3308558.3313565. URL https://dl.acm.org/doi/10.1145/3308558.3313565.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In *Advances in neural information processing systems*, volume 28, pp. 505–513, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/bac9162b47c56fc8a4d2a519803d51b3-Abstract.html.

Paul Pimsleur. A memory schedule. *The Modern Language Journal*, 51(2):73–75, 1967. doi: 10.1111/j.1540-4781.1967.tb06700.x.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15*, pp. 593–607. Springer, 2018. doi: 10.1007/978-3-319-93417-4_38. URL https://arxiv.org/abs/1703.06103.

Burr Settles and Brendan Meeder. A trainable spaced repetition model for language learning. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 1848–1858, 2016. doi: 10.18653/v1/P16-1174. URL https://aclanthology.org/P16-1174/.

Mingxing Shao, Tiancheng Zhang, Minghe Yu, Zhenghao Liu, Yifang Yin, Hengyu Liu, and Ge Yu. Leveraging student profiles and the mamba framework to enhance knowledge tracing. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pp. 201–218, 2025. doi: 10.1007/978-3-032-06109-6_12.

Xiaoxuan Shen, Fenghua Yu, Yaqi Liu, Ruxia Liang, Qian Wan, Kai Yang, and Jianwen Sun. Revisiting knowledge tracing: A simple and powerful model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 263–272, 2024. doi: 10.1145/3664647.3681205.

Xiaoxuan Shen, Zhihai Hu, Di Chen, Jianwen Sun, and Shengyingjie Liu. ROKAN: Toward interpretable and domain-robust memory behavior modeling. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 2588–2598, 2025a. doi: 10.1145/3746252.3761087.

Xiaoxuan Shen, Fenghua Yu, Qian Wan, Ruxia Liang, and Jianwen Sun. Multi-level contrastive learning for knowledge tracing. *ACM Transactions on Knowledge Discovery from Data*, 19(8): 1–29, 2025b. doi: 10.1145/3759920.

Xiaoxuan Shen, Zhihai Hu, Qirong Chen, and Pei Wang. Evolvable psychology informed neural network for memory behavior modeling. *Information Processing & Management*, 63(1):104312, 2026a. doi: 10.1016/j.ipm.2025.104312.

Xiaoxuan Shen, Zhihai Hu, Fuqing Li, Shengyingjie Liu, and Jianwen Sun. Invariant representation learning for memory behavior modeling via adaptive environment separation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2047–2055, 2026b. doi: 10.1609/aaai.v40i3.37186.

Jingyong Su, Junyao Ye, Liqiang Nie, Yilong Cao, and Yongyong Chen. Optimizing spaced repetition schedule by capturing the dynamics of memory. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10085–10097, 2023. doi: 10.1109/TKDE.2023.3251721.

Jianwen Sun, Fenghua Yu, Qian Wan, Qing Li, Sannyuya Liu, and Xiaoxuan Shen. Interpretable knowledge tracing with multiscale state representation. In *Proceedings of the ACM Web Conference 2024*, pp. 3265–3276, 2024. doi: 10.1145/3589334.3645373.

Jianwen Sun, Qirong Chen, Zhenya Huang, Zhihai Hu, Ruxia Liang, and Xiaoxuan Shen. Combining denoised neural network and genetic symbolic regression for memory behavior modeling via dynamic asynchronous optimization. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2735–2746, 2025. doi: 10.1145/3711896.3736886.

Behzad Tabibian, Utkarsh Upadhyay, Abir De, Ali Zarezade, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Enhancing human learning via spaced repetition optimization. *Proceedings of the National Academy of Sciences*, 116(10):3988–3993, 2019.

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*, volume 30, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Qinfeng Xiao and Jing Wang. Drl-srs: A deep reinforcement learning approach for optimizing spaced repetition scheduling. *Applied Sciences*, 14(13):5591, 2024. doi: 10.3390/app14135591.

Jia Xu, Rongrong Tang, Pin Lv, Minghe Yu, Ge Yu, and Enhong Chen. LTKT: Knowledge tracing based on positive and negative learning transfers. *Tsinghua Science and Technology*, 31(3): 1894–1917, 2026. doi: 10.26599/TST.2024.9010201.

Junyao Ye, Jingyong Su, and Yilong Cao. A stochastic shortest path algorithm for optimizing spaced repetition scheduling. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4381–4390, 2022. doi: 10.1145/3534678.3539081.

# APPENDIX CONTENTS


| App. | Section and detailed contents | Page |
|---|---|---|
| A | **Detailed Related Work.** Memory and spaced-repetition models; sequence and forgetting-aware knowledge tracing; relational modeling; capacity-aware scheduling. | 14 |
| I | **Model-Agnostic Effect of Open-Space Retrieval.** Test H-SMAPE of all 27 baselines with and without the shared retrieval layer under the source-only Row protocol. | 24 |
| J | **Discussion, Limitations, and Future Work.** Supported prediction and scheduling claims; dataset domains, log granularity, label support, and simulation horizon; prospective and cross-platform evaluation. | 24 |
| B | **Core Ablation and Decision-Case Details.** Core-factorial design and inference; exact-known versus exact-unknown attribution; early/late-review cases and counterexample. | 14 |
| C | **Supplementary Confirmation on Held-out Anki Users.** Cohort provenance; identity-disjoint and chronological protocols; execution audit; six-model and 27-baseline results; estimand and uncertainty limits. | 16 |
| D | **Complete Source-Only Generalization Results.** Row/Learner/Item/State/Joint construction; source-only information boundary; full 27-baseline H-SMAPE and P-RMSE reporting. | 17 |
| E | **Feature Specification and Construction.** Event features; masked BiGRU and outcome-aware fusion; static and multiscale transition features; strict- isolation feature boundary. | 18 |
| F | **Evaluation, Data, and Training Details.** Label-access boundary and four capability controls; data/label and partition construction; optimization and implementation; baseline adaptations. | 20 |
| G | **Fixed-Scheduler Evaluation Protocols.** OSR-ACRI daily loop; next-review decision metrics; 365-day fixed-scheduler simulation; fixed-predictor scheduler evaluation. | 21 |
| H | **Metric Definitions and Reporting.** Statistical units and estimands; half-life and recall-probability metric formulas; complete metric tables for all datasets and capability-control views. | 23 |

# A Detailed Related Work

**Memory and spaced-repetition models.** HLR learns log half-life from practice statistics (Settles & Meeder, 2016); DASH adds difficulty and spacing-sensitive history (Lindsey et al., 2014); ACT-R uses power-law activation (Anderson et al., 1997); and DHP and DSR-style controls define structured memory-state updates (Ye et al., 2022; Su et al., 2023). These methods are attractive because their state variables have a direct scheduling interpretation: a half-life or stability estimate can be converted into a target-recall interval. They also illustrate the central trade-off in this domain. Strong hand-designed dynamics are interpretable, but they can make the effect of a recall or forgetting event too rigid; flexible predictors can fit heterogeneous histories, but their outputs are harder to audit when they are used for review decisions. Recent psychology-informed models combine neural sequence representations with evolved symbolic descriptors (Shen et al., 2026a; Sun et al., 2025), while THLR uses a Transformer to estimate memory trajectories (Xiao & Wang, 2024). Our primary task remains observed half-life estimation rather than learning a new cognitive theory. The difference is that OSR-HLR predicts relative half-life, aggregates repeated states, and applies protocol-restricted residual corrections whose sources remain visible. We additionally use history-reencoded long-horizon simulation as a policy stress test, with Validation-fitted transition environments and no claim of observed policy outcomes.

**Sequence models of learning and forgetting.** Recurrent networks and knowledge tracing retain ordered interaction histories (Cho et al., 2014; Piech et al., 2015; Nagatani et al., 2019). Forgetting-aware models add elapsed-time features or attention (Ma et al., 2023; Bai et al., 2025; Guan et al., 2025). Recent comparators also include simple response-update units (Shen et al., 2024), multiscale and contrastive state representations (Sun et al., 2024; Shen et al., 2025b), and state-space sequence models (Shao et al., 2025). These architectures are well suited to ordered histories, and we include versions of this family in the baseline suite. However, many knowledge-tracing formulations are optimized for next-response probability, whereas the present application needs a calibrated half-life that can be used to schedule an interval. OSR-HLR therefore keeps sequence modeling inside a half-life regression interface. The sequence representation retains outcome information used in state keys. The core ablations evaluate relative targeting and residual transfer as distinct mechanisms within the complete prediction system.

**Relational modeling.** Graph models can encode typed dependencies among entities (Schlichtkrull et al., 2018; Hu et al., 2020). Knowledge-tracing graphs have represented concept influence and positive or negative learning transfer (Cui et al., 2024; Xu et al., 2026). Our residual graph is deliberately simpler than a learned GNN. Its nodes are recurring memory states, its relations are deterministic, and its messages are clipped visible-label residuals. This design trades representation flexibility for a clear information boundary. At Test time, every correction can be traced to Train/Validation states allowed by the evaluation protocol, and the current held-out outcome is never available to the graph. The graph is therefore closer to a controlled evidence-transfer layer than to an end-to-end relational encoder. This distinction matters for the source-only generalization protocols: exact state reuse, parent-state evidence, temporal neighbors, and fallback relations can be enabled or disabled without changing the neural prediction path.

**Capacity-aware scheduling.** Threshold rules convert a predicted half-life into a target-recall interval, while DHP, SSP-MMC, MEMORIZE (Tabibian et al., 2019), and Anki-style policies encode alternative memory or review objectives (Ye et al., 2022; Su et al., 2023). These methods do not jointly address online predictor calibration and competition among simultaneously due items under a hard daily capacity. A threshold rule can be reasonable when capacity is unconstrained, but once many cards are overdue the scheduler must decide which delay is most harmful, not merely which item has crossed a recall threshold. OSR-ACRI addresses this scheduling problem with three ingredients: one-sided calibration after overconfident outcomes, marginal-delay ranking among due items, and queue-age terms that reduce long-tail starvation. The fixed-predictor evaluation is designed around this distinction. It holds OSR-HLR predictions, simulated environments, initial states, random streams, and workload constant, then compares schedulers under the same source-domain simulator. This supports a scheduler-level offline result, while prospective learner benefit remains outside the retrospective evidence.

# B Core Ablation and Decision-Case Details

**Core-factorial design.** The core analysis contains four cells: the complete method, direct $\log h$ prediction with the residual graph retained, relative $\log(h/\Delta t)$ prediction with only exact-state retrieval, and the joint removal of both mechanisms. All cells use the source-only protocols of the main comparison: labels are refitted within each split, residual evidence comes only from Train and Validation states, and historical-strength and state-multiplicity inputs are disabled. MaiMemo and Anki use the Row protocol with seeds 42–51, and EdNet-Recall uses all five source-only tracks with seeds 42–44. For each dataset and seed we train one relative-target and one direct-target model with the same command, inputs, hidden width, optimizer, and loss; only the target parameterization differs. The graph and exact-only cells are two evaluation views of the same checkpoint, so

removing the residual graph never changes the trained model. Checkpoints are selected by Validation H-SMAPE; Test targets are not used for training or selection. The joint cell completes the target-parameterization by residual-transfer factorial rather than treating the encoder alone as the source of the gain.

**Inference and decision-case audit.** Ablation effects are defined as ablated-minus-OSR-HLR Test H-SMAPE, so positive values mean that removal hurts. We report paired seed or track–seed differences and paired 95% intervals for the nine dataset–core-ablation comparisons; native-SRS comparisons additionally report two-sided paired $t$-tests, all with $p \leq 0.002$. The two-by-two interaction is also recorded. It is $+0.386$ points (95% CI $[+0.312, +0.460]$) on MaiMemo and $+0.829$ points ($[+0.477, +1.180]$) on Anki. Across the five EdNet-Recall tracks it is $+0.316$ points ($[-0.115, +0.746]$). The native-SRS interactions show that the direct-target penalty is larger without residual transfer, so the two mechanisms are complementary. The decision cases use the learner/item/state-disjoint MaiMemo Test set and $q = 0.90$. Early- and late-review corrections are selected nearest the median baseline absolute log error within the corresponding error direction; the counterexample is nearest the median increase in OSR-HLR error. Ties use the smallest original node identifier. Because this case selection uses observed targets, it is descriptive and does not replace aggregate inference. EdNet-Recall is included in the predictive mechanism inference but not in this scheduling illustration because its adaptive TOEIC log has no native spaced-repetition due-date semantics.

**Exact-known and exact-unknown state attribution.** This diagnostic uses the transductive label-pool view on the full released tables (Appendix F), where released labels are shared by all observations of a state; it is not part of the headline source-only comparison. We decompose Test states according to whether their complete state key is available among protocol-visible historical states. This separates exact-state reuse from prediction on previously unseen states. We report the gain on exact-unknown states and the share of within-method correction attributable to exact lookup.

Table 4 reports this decomposition in separate column groups. Let $K$ denote exact-known Test observations whose complete state key appeared in Train or Validation, and let $U$ denote exact-unknown observations whose complete state key did not. For recurrence prevalence $\pi$, write $R_m(\pi) = \pi R_{m,K} + (1-\pi)R_{m,U}$. The middle columns report baseline and OSR-HLR errors on $U$. The rightmost column reports the share of OSR-HLR's base-to-final improvement attributable to exact lookup. EdNet-Recall has zero exact-known Test observations by construction because its state identity includes the split partition.

Table 4: Exact-known and exact-unknown state attribution. The center columns report baseline and OSR-HLR errors on exact-unknown states; the rightmost column reports the within-OSR-HLR exact-lookup share.

| Dataset | Exact-known | **Exact-unknown states $U$** | | | **Within OSR-HLR** |
|---|---|---|---|---|---|
| | | Baseline error | OSR-HLR error | Gain [95% CI] | Exact-lookup share |
| MaiMemo | 9.4% | 15.428 | 13.439 | **1.988** [1.906, 2.071] | 94.4% |
| Anki | 69.0% | 22.755 | 14.438 | **8.317** [7.544, 9.091] | 96.4% |
| EdNet-Recall | 0.0% | 61.717 | 34.098 | **27.620** [25.188, 30.051] | 0.0% |

Exact lookup accounts for 94.4% and 96.4% of the within-method improvement on MaiMemo and Anki, where recurring Test states are common. On exact-unknown states, OSR-HLR improves H-SMAPE over the baseline by 1.988, 8.317, and 27.620 points on MaiMemo, Anki, and EdNet-Recall, respectively. Together, these results characterize OSR-HLR as a query-conditioned prediction system combining exact historical reuse with relational residual transfer for previously unseen states.

**Decision examples.** Figure 9 traces the full prediction-to-decision chain for three deterministic source-only Test cases selected after the aggregate analysis: an early-review correction, a late-review correction, and a counterexample. The horizontal coordinate is the recommended interval relative to the independently fitted ideal interval, and the vertical coordinate is the realized recall probability minus the target $q = 0.90$. Points near the origin are therefore both well timed and well calibrated for the review decision.

For easy card #10792, the reference half-life is 1058.6 days, giving an ideal $q = 0.90$ interval of 160.9 days. The baseline predicts 885.6 days and reviews after 134.6 days, 16.3% early, reaching 91.6% recall and overshooting the target. OSR-HLR predicts 1093.2 days and reviews after 166.2 days, only 3.3% late, reaching 89.7% recall; the target-recall error falls by 80.2%. For hard card #12223, the reference half-life is 147.6 days and the ideal interval is 22.4 days. The baseline predicts 175.3 days, reviews 18.8% late, and falls to 88.2% recall. OSR-HLR predicts 154.5 days, reviews 4.7% late, and reaches 89.6%, reducing target-recall error by 75.0%. The third panel keeps a counterexample: for card #11503 the baseline is nearly ideal, while OSR-HLR reviews later and increases target-recall error. Across all 2,793 source-only Test observations, OSR-HLR has lower absolute log-half-life error on 61.0%; the figure is a mechanistic illustration rather than a separate statistical test.

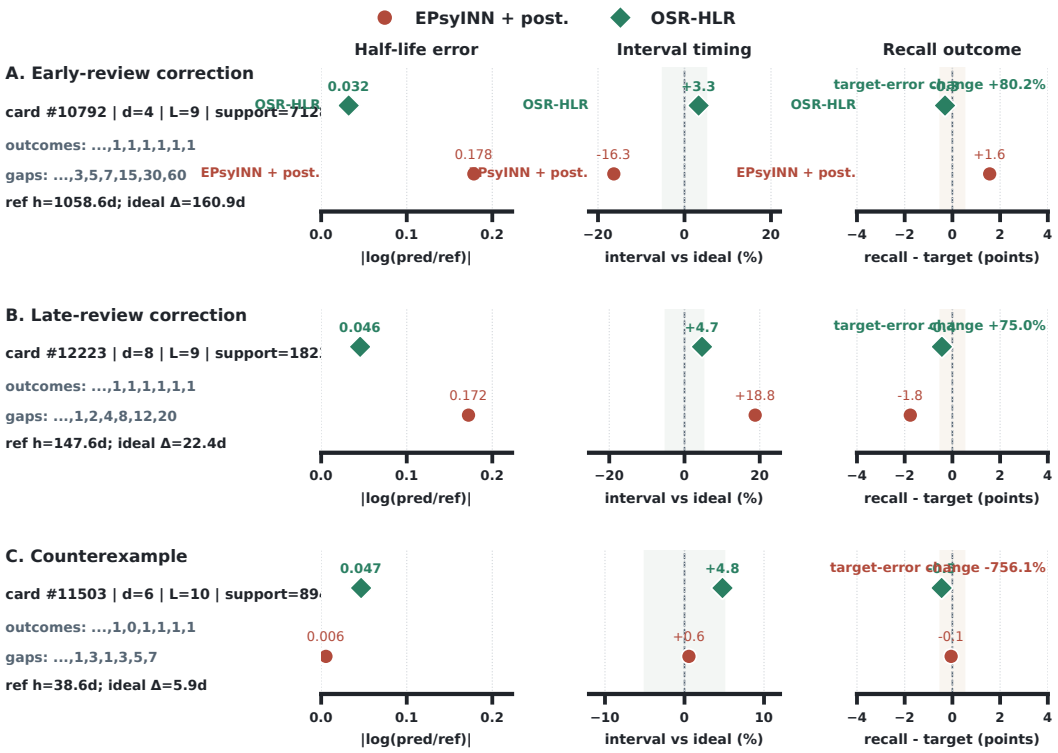


Figure 9: Decision examples at $q = 0.90$. Rows are deterministic source-only Test cases; columns show the history context, absolute log-half-life error, review-interval error relative to the ideal interval, and recall error relative to the 90% target.

## C Supplementary Confirmation on Held-out Anki Users

**Held-out user provenance and lock.** The source is the gated Anki Revlogs 10K release. A salted manifest fixed 1,000 of the 10,000 user collections for development before the confirmation data were downloaded or inspected. The confirmatory set was defined as every remaining user identifier after excluding two schema-pilot collections, yielding 8,998 prespecified collections. One collection could not form any valid repeated-review history, so 8,997 collections entered the experiment.

**Identity and chronological protocols.** Identity evaluation assigns learner, collection-scoped card, and complete state independently and retains only observations for which all three assignments agree; Train, Validation, and Test therefore have zero overlap on these identities. Chronological evaluation instead uses each learner's relative lifetime and within-day order, with the earliest 60% of records for Train, the next 20% for Validation, and the final 20% for Test. This evaluates future prediction for observed learners rather than new-learner transfer. Table 5 reports the resulting partition sizes.

Table 5: Held-out Anki users: data partitions, giving Train/Validation/Test observations and states for the identity-disjoint and within-learner chronological protocols.

| | Train | | Validation | | Test | |
|---|---|---|---|---|---|---|
| Protocol | Observations | States | Observations | States | Observations | States |
| Identity-disjoint | 63,175 | 6,863 | 20,908 | 2,423 | 23,242 | 2,501 |
| Chronological | 464,330 | 52,440 | 179,686 | 19,904 | 183,471 | 19,125 |

**Information boundary and execution audit.** For both protocols, labels are refitted within each partition, each Test observation is an independent target node, and only Train/Validation states can provide residual labels. Historical-strength and state-multiplicity inputs are disabled. Mutating current Test half-life or recall labels changes no prediction. An execution audit found an observation-order mismatch in the initial chronological output before aggregate results were released; all 60 affected runs were invalidated and retrained after key-based alignment. The final audit reports zero observation mismatch and zero source–Test state overlap in both protocols.

Table 6: Held-out Anki users: source-only Test H-SMAPE and relative improvement for all 27 baselines and OSR-HLR under identity-disjoint and within-learner chronological evaluation.

| Method | Identity-disjoint | | Chronological | |
|---|---|---|---|---|
| | H-SMAPE ↓ | Improvement (%) ↑ | H-SMAPE ↓ | Improvement (%) ↑ |
| *Cognitive / formulaic* | | | | |
| Pimsleur | $99.234 \pm 0.017$ | **+62.48%** | $78.888 \pm 0.008$ | **+43.68%** |
| Leitner | $109.813 \pm 0.009$ | **+66.10%** | $99.538 \pm 0.102$ | **+55.37%** |
| ACT-R | $77.430 \pm 0.019$ | **+51.92%** | $73.208 \pm 0.062$ | **+39.31%** |
| HLR | $72.937 \pm 0.027$ | **+48.96%** | $68.884 \pm 0.049$ | **+35.50%** |
| DASH | $70.607 \pm 0.119$ | **+47.27%** | $67.152 \pm 0.056$ | **+33.84%** |
| DHP | $78.759 \pm 0.157$ | **+52.73%** | $72.998 \pm 0.510$ | **+39.14%** |
| FSRS | $55.855 \pm 4.522$ | **+33.35%** | $56.045 \pm 0.304$ | **+20.73%** |
| *Generic sequence* | | | | |
| RNN | $43.628 \pm 0.500$ | **+14.67%** | $46.665 \pm 0.195$ | **+4.79%** |
| GRU | $43.688 \pm 0.864$ | **+14.78%** | $46.722 \pm 0.227$ | **+4.91%** |
| LSTM | $43.476 \pm 0.550$ | **+14.37%** | $46.707 \pm 0.147$ | **+4.88%** |
| Transformer | $45.023 \pm 0.674$ | **+17.31%** | $46.448 \pm 0.598$ | **+4.35%** |
| DKT | $43.135 \pm 1.082$ | **+13.69%** | $46.421 \pm 0.297$ | **+4.29%** |
| *Memory / forgetting* | | | | |
| FIFAKT | $42.034 \pm 1.257$ | **+11.43%** | $46.478 \pm 0.412$ | **+4.41%** |
| GRU-HLR | $42.973 \pm 0.920$ | **+13.37%** | $46.375 \pm 0.330$ | **+4.20%** |
| THLR | $44.408 \pm 0.959$ | **+16.17%** | $47.000 \pm 0.169$ | **+5.47%** |
| ROKAN | $56.808 \pm 1.391$ | **+34.46%** | $50.302 \pm 0.556$ | **+11.68%** |
| I-Mem | $43.925 \pm 0.814$ | **+15.24%** | $46.377 \pm 0.281$ | **+4.20%** |
| *Recent KT / memory* | | | | |
| EPsyINN | $43.741 \pm 0.393$ | **+14.89%** | $46.310 \pm 0.318$ | **+4.06%** |
| LefoKT | $46.053 \pm 0.890$ | **+19.16%** | $45.521 \pm 0.178$ | **+2.40%** |
| ReKT | $41.999 \pm 0.694$ | **+11.36%** | $46.699 \pm 0.348$ | **+4.86%** |
| LTKT | $43.599 \pm 0.709$ | **+14.61%** | $46.322 \pm 0.149$ | **+4.09%** |
| SPsyINN | $44.426 \pm 0.388$ | **+16.20%** | $46.483 \pm 0.227$ | **+4.42%** |
| KVFKT | $42.859 \pm 0.334$ | **+13.14%** | $45.925 \pm 0.319$ | **+3.26%** |
| GRKT | $43.324 \pm 0.392$ | **+14.07%** | $46.676 \pm 0.295$ | **+4.82%** |
| MIKT | $43.078 \pm 0.350$ | **+13.58%** | $46.793 \pm 0.188$ | **+5.05%** |
| MCKT | $43.940 \pm 0.918$ | **+15.27%** | $46.805 \pm 0.235$ | **+5.08%** |
| DGMKT | $43.589 \pm 0.601$ | **+14.59%** | $46.242 \pm 0.327$ | **+3.92%** |
| *Ours* | | | | |
| **OSR-HLR (Ours)** | **$37.229 \pm 0.307$** | – | **$44.428 \pm 0.310$** | – |

**Models and primary inference.** The original six-model, ten-seed family compares OSR-HLR with FSRS, DHP, GRU, Transformer, and GRU-HLR under the same source-only postprocessor. OSR-HLR reduces H-SMAPE relative to GRU-HLR by **13.37%** under identity isolation and **4.20%** under chronological evaluation. Primary inference uses paired seed differences, paired 95% intervals, and exact sign tests across the two protocols. Additional half-life and recall-probability metrics are reported as a post-hoc OSR-HLR–versus–GRU-HLR robustness audit; the complete 27-baseline breadth extension below is available for source-only H-SMAPE.

For breadth, we separately froze a post-hoc extension covering all 27 baselines and OSR-HLR. Every method uses the same split, ten seeds, source-only Test inference, and deterministic residual postprocessor. Table 6 contains all 2 protocols $\times$ 28 models $\times$ 10 seeds $=$ 560 model–protocol–seed units; 340 units are hash-linked reuse and 220 are newly trained chronological baseline runs.

**Estimand and uncertainty limits.** The identity gain remains positive under paired-seed, equal-state, equal-learner, and label-refit uncertainty analyses. The chronological observation/state/label estimates are positive, but its equal-learner interval includes zero. Observation-weighted recall metrics also reverse for the identity protocol. The confirmation therefore supports observation- and state-level prediction on previously unused Anki users, not uniform per-learner benefit or transfer to a different platform.

## D Complete Source-Only Generalization Protocol Results

**Protocol construction.** The row protocol protects no identity; learner, item, and state protocols protect their named key; and the joint protocol retains only observations whose independent learner, item, and complete-state assignments agree. Within the native-SRS datasets, the other four protocols are deterministically subsampled to the joint protocol's Train/Validation/Test observation counts without using targets: 7,348/2,802/2,833 for MaiMemo and 8,511/2,089/1,631 for Anki. Anki item identity is collection-scoped, so learner isolation also separates those card identities. EdNet-Recall uses the same five estimands with global question identifiers: Row keeps all qualified delayed-response rows, Learner removes learner overlap, Item removes global-question overlap, State removes complete memory state overlap, and Joint removes all three.

**Training and information boundary.** All 27 baselines and OSR-HLR are trained in every dataset–protocol cell: seeds 42–51 for MaiMemo/Anki and seeds 42–44 for EdNet-Recall. Every method uses the same

deterministic source-only postprocessor; each Test observation is a separate target node, only Train/Validation states provide residual labels, and historical-strength and state-multiplicity features are disabled for the native-SRS source-only protocols. Checkpoints are selected on Validation only. This produces 3,220 model runs in the reported source-only generalization suite (2,800 native-SRS runs and 420 EdNet-Recall runs); run counts describe coverage, not independent statistical evidence.

**Reporting and inference.** Table 7 reports H-SMAPE for all 27 baselines and OSR-HLR across the five source-only generalization protocols. The Improvement entry compares OSR-HLR with the strongest of all 27 baselines in each dataset and protocol. P-RMSE appears in Figure 10.

The MaiMemo H-SMAPE reductions are **17.40%**, **16.18%**, **15.49%**, **14.35%**, and **18.55%** for row, learner, item, state, and joint isolation; the corresponding Anki reductions are **2.59%**, **2.78%**, **1.99%**, **6.40%**, and **1.54%**; the EdNet-Recall reductions are **42.46%**, **36.71%**, **14.05%**, **29.66%**, and **22.94%**. Statistical support is strongest for all five MaiMemo results and the Anki learner- and state-disjoint results; Anki row, item, and joint are weaker. P-RMSE ranks first in all ten native-SRS settings, with reductions of **15.51–20.33%** on MaiMemo and **4.06–9.34%** on Anki. On EdNet-Recall, P-RMSE reductions are **1.18–7.61%**. These protocols define different held-out estimands, so their gains are not additive.

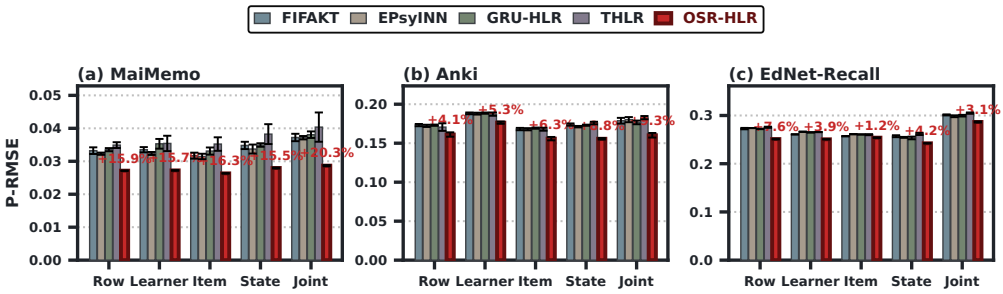


Figure 10: Source-only protocols: P-RMSE with representative baselines.

Table 7: Five source-only generalization protocols: Test H-SMAPE for all baselines and OSR-HLR; lower is better. **Red bold**, blue underlined, and orange mark ranks 1–3 in each column.

| Method | Row | | | Learner | | | Item | | | State | | | Joint | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MaiMemo | Anki | EdNet | MaiMemo | Anki | EdNet | MaiMemo | Anki | EdNet | MaiMemo | Anki | EdNet | MaiMemo | Anki | EdNet |
| *Cognitive / formulaic* | | | | | | | | | | | | | | | |
| Pimsleur | 63.79 | 100.02 | 135.46 | 64.31 | 106.45 | 147.33 | 63.99 | 99.63 | 148.79 | 66.93 | 112.00 | 153.97 | 61.54 | 106.24 | 184.11 |
| Leitner | 61.15 | 94.93 | 137.52 | 60.36 | 102.13 | 151.84 | 61.13 | 95.85 | 153.69 | 61.39 | 107.42 | 160.55 | 61.56 | 103.32 | 176.68 |
| ACT-R | 51.70 | 73.73 | 77.52 | 52.06 | 77.63 | 68.08 | 51.71 | 71.95 | 62.42 | 52.61 | 79.94 | 65.87 | 53.07 | 77.13 | 70.81 |
| HLR | 50.29 | 70.44 | 74.04 | 50.50 | 76.46 | 62.46 | 50.24 | 71.40 | 49.93 | 52.56 | 80.15 | 73.74 | 48.53 | 72.13 | 134.18 |
| DASH | 42.29 | 70.05 | 75.67 | 41.84 | 76.02 | 63.56 | 41.73 | 70.86 | 49.55 | 42.98 | 79.47 | 62.93 | 39.46 | 71.08 | 154.71 |
| DHP | 47.65 | 72.06 | 82.94 | 47.31 | 76.94 | 66.97 | 47.05 | 72.01 | 56.64 | 49.14 | 79.40 | 68.09 | 49.68 | 89.26 | 109.83 |
| FSRS | 32.63 | 52.25 | 69.83 | 32.78 | 60.43 | 55.98 | 32.20 | 52.42 | 42.46 | 35.76 | 62.13 | 52.00 | 32.23 | 64.62 | 110.18 |
| *Generic sequence* | | | | | | | | | | | | | | | |
| RNN | 17.56 | 40.77 | 60.46 | 17.27 | 53.10 | 53.08 | 17.09 | 40.32 | 38.41 | 16.67 | 47.96 | 48.00 | 18.22 | 52.31 | 63.63 |
| GRU | 16.81 | 41.21 | 72.05 | 17.19 | 53.26 | 52.32 | 16.63 | 40.54 | 38.12 | 16.46 | 48.41 | 51.85 | 17.86 | 53.57 | 67.46 |
| LSTM | 17.03 | 40.91 | 69.30 | 17.14 | 52.88 | 52.40 | 16.68 | 40.22 | 38.61 | 16.59 | 47.98 | 53.57 | 17.95 | 53.18 | 62.10 |
| Transformer | 17.55 | 42.17 | 70.59 | 17.89 | 54.21 | 55.07 | 17.40 | 41.03 | 41.57 | 18.20 | 50.15 | 60.34 | 19.28 | 54.58 | 63.26 |
| DKT | 16.61 | 40.12 | 61.40 | 16.66 | 51.80 | 52.06 | 16.18 | 39.23 | 34.13 | 16.03 | 46.77 | 54.13 | 17.62 | 50.31 | 62.88 |
| *Memory / forgetting* | | | | | | | | | | | | | | | |
| FIFAKT | 16.80 | 39.34 | 62.17 | 16.80 | 50.91 | 47.15 | 16.28 | 38.10 | 32.86 | 16.41 | 47.97 | 54.73 | 17.41 | 53.63 | 60.41 |
| GRU-HLR | 16.68 | 40.88 | 59.26 | 17.32 | 52.99 | 51.82 | 16.68 | 40.11 | 38.24 | 16.69 | 46.68 | 48.56 | 17.72 | 52.83 | 63.31 |
| THLR | 17.51 | 42.86 | 62.72 | 17.74 | 53.16 | 52.08 | 17.91 | 41.01 | 36.91 | 18.14 | 50.27 | 55.17 | 18.58 | 56.26 | 62.52 |
| ROKAN | 31.38 | 47.48 | 64.31 | 31.28 | 60.92 | 55.70 | 31.35 | 47.34 | 43.82 | 26.56 | 62.13 | 69.64 | 31.10 | 71.13 | 71.12 |
| I-Mem | 18.94 | 40.96 | 60.73 | 18.17 | 52.18 | 53.24 | 18.04 | 40.29 | 39.14 | 18.34 | 47.79 | 57.99 | 18.84 | 54.19 | 75.79 |
| *Recent KT / memory* | | | | | | | | | | | | | | | |
| EPsyINN | 16.28 | 40.33 | 61.24 | 16.35 | 52.63 | 53.23 | 16.10 | 39.29 | 39.16 | 16.32 | 46.49 | 46.51 | 17.35 | 53.77 | 63.51 |
| LefoKT | 17.66 | 42.34 | 66.91 | 18.68 | 56.88 | 54.63 | 18.03 | 43.06 | 39.54 | 17.98 | 50.73 | 55.70 | 19.03 | 60.44 | 73.20 |
| ReKT | 17.09 | 41.53 | 61.64 | 17.03 | 52.21 | 53.32 | 16.84 | 39.73 | 39.04 | 16.62 | 46.42 | 46.88 | 17.90 | 52.45 | 63.68 |
| LTKT | 16.26 | 40.74 | 62.18 | 16.16 | 53.77 | 52.76 | 15.90 | 39.55 | 39.13 | 16.45 | 46.46 | 46.20 | 17.37 | 55.01 | 66.85 |
| SPsyINN | 16.68 | 39.74 | 59.99 | 16.34 | 53.22 | 53.06 | 16.13 | 38.81 | 38.88 | 16.48 | 46.64 | 46.63 | 17.36 | 54.50 | 66.67 |
| KVFKT | 17.35 | 43.63 | 66.69 | 17.20 | 54.41 | 53.26 | 17.19 | 42.74 | 38.92 | 17.09 | 50.28 | 47.60 | 18.02 | 61.60 | 65.77 |
| GRKT | 17.71 | 42.07 | 71.96 | 17.78 | 54.89 | 53.24 | 17.35 | 41.57 | 39.16 | 17.41 | 49.08 | 46.91 | 18.68 | 57.33 | 67.59 |
| MIKT | 16.68 | 41.70 | 62.32 | 16.70 | 53.43 | 53.27 | 16.19 | 40.56 | 39.05 | 16.28 | 47.65 | 46.59 | 17.54 | 55.41 | 60.57 |
| MCKT | 16.47 | 40.35 | 60.83 | 16.49 | 53.04 | 52.22 | 16.10 | 39.26 | 38.76 | 16.09 | 47.00 | 47.54 | 17.37 | 53.16 | 64.94 |
| DGMKT | 16.83 | 39.38 | 60.99 | 16.65 | 52.46 | 53.35 | 16.36 | 38.48 | 39.18 | 16.40 | 46.48 | 47.99 | 17.52 | 50.98 | 65.19 |
| *Ours* | | | | | | | | | | | | | | | |
| **OSR-HLR (Ours)** | **13.43** | **38.32** | **34.10** | **13.54** | **49.49** | **29.84** | **13.44** | **37.34** | **28.25** | **13.73** | **43.45** | **32.49** | **14.13** | **49.54** | **46.55** |
| **Improvement (%)** | **+17.40%** | **+2.59%** | **+42.46%** | **+16.18%** | **+2.78%** | **+36.71%** | **+15.49%** | **+1.99%** | **+14.05%** | **+14.35%** | **+6.40%** | **+29.66%** | **+18.55%** | **+1.54%** | **+22.94%** |

# E Feature Specification and Construction

This appendix instantiates the abstract inputs $\mathbf{X}_i \in \mathbb{R}^{T_{\max} \times F_e}$, with valid length $L_i \le T_{\max}$, and $\mathbf{s}_i \in \mathbb{R}^{F_s}$ used in Section 4. All features are computed from the current query and its preceding review history; the current recall or half-life target is never an input. Here $T_{\max} = 44$, $F_e = 11$, and $F_s = 101$.

## E.1 Event features

The most recent $T_{\max}$ events are kept in chronological order and right-padded under a validity mask. Let $a_k$ be the raw rating, $s_k \in \{-1, +1\}$ its forget/recall sign, $\ell_k = \log(1 + t_k)$, $\eta_k$ the normalized historical parent

strength, and $A_k$ its availability. Each event has

$$\mathbf{x}_{ik} = [\, a_k/4,\ s_k,\ \ell_k/8,\ (\ell_k - \ell_{k-1})/8,\ k/T_{\max},\\ \eta_k,\ A_k,\ z_k^{\text{lag}},\ s_k z_k^{\text{lag}},\ z_k^{\text{align}},\ s_k z_k^{\text{align}}\,],$$

where

$$z(t,h) = \text{clip}\left(\frac{\log(1 + t/\max(h, 10^{-4}))}{8}, 0, 5\right).$$

The lagged and aligned variants use the previous and same-event historical strength, respectively. When strength is unavailable, its value and derived spacing channels are zero and $A_k = 0$ distinguishes missingness from a small measured value. Table 8 lists the 11 event channels.

Table 8: The 11 features for each of at most 44 review events.

| Index | Feature |
|---|---|
| 1 | raw rating divided by 4 |
| 2 | recall/forget sign |
| 3 | $\log(1 + \text{interval})/8$ |
| 4 | adjacent log-interval difference |
| 5 | position divided by 44 |
| 6 | $\log(1 + \text{historical parent strength})/8$ |
| 7 | parent-strength availability mask |
| 8 | one-step-lagged relative spacing |
| 9 | outcome-signed lagged spacing |
| 10 | parent-aligned spacing |
| 11 | outcome-signed parent-aligned spacing |

### E.2 Sequence encoder and fusion

Each event is projected to a learned representation and encoded by a masked bidirectional GRU:

$$\mathbf{e}_{ik} = \text{GELU}(\text{LN}(\mathbf{W}_e \mathbf{x}_{ik} + \mathbf{b}_e)), \qquad \mathbf{g}_{i1:L_i} = \text{BiGRU}(\mathbf{e}_{i1:L_i}). \tag{11}$$

In addition to a learned content score, attention receives an outcome-aware basis $\boldsymbol{\phi}_{ik} = \phi(\mathbf{X}_i, k)$ describing recency, outcome transitions, streaks, relative spacing, the last forgetting event, and the subsequent recovery segment:

$$\tilde{a}_{ik} = \mathbf{w}_a^\top \mathbf{g}_{ik} + \boldsymbol{\beta}^\top \boldsymbol{\phi}_{ik}, \qquad \alpha_{ik} = \frac{m_{ik} \exp(\tilde{a}_{ik})}{\sum_\ell m_{i\ell} \exp(\tilde{a}_{i\ell})}. \tag{12}$$

Depth-dependent gates suppress complex history terms when the sequence is too short to support them. The attention summary, terminal forward/backward states, projected static vector, and difficulty embedding are fused as

$$\mathbf{z}_i = [\textstyle\sum_k \alpha_{ik} \mathbf{g}_{ik}; \mathbf{g}_i^{\text{final}}; \psi_s(\mathbf{s}_i); \text{Emb}(d_i)]. \tag{13}$$

A main prediction head and bounded local correction heads map $\mathbf{z}_i$ to $o_i$. The corrections specialize recent-prefix, singleton-state, and history-depth behavior without changing the shared encoder.

### E.3 Static and transition features

Let $c_g$ be group support, $c_i$ the review count underlying observation $i$, $n_v$ the number of observations sharing state $v$, and $L_i$ the observed history length. The nine query/count features are

$$\mathbf{q}_i = [\, \log(1 + \Delta t_i)/8,\ \sqrt{\Delta t_i}/32,\ (1 + \Delta t_i)^{-1},\\ \log(1 + c_g)/16,\ \log(1 + c_i)/16,\ \min(c_i/c_g, 1),\\ \log(1 + |c_g - c_i|)/16,\ L_i/T_{\max},\ \log(1 + n_v)/4\,].$$

For each window in $\mathcal{W} = \{4, 8, 16, 32, 44\}$, the model computes 16 statistics: RR/RF/FR/FF transition proportions, switch proportion, last sign, current streak, longest recall and forget streaks, recall proportion, recall and forget mean log intervals and their difference, signed interval mean, outcome–interval covariance, and interval slope. These give 80 dimensions. Twelve full-history features add distances to the latest recall and forget, the two longest streaks, and outcome/interval decay summaries with powers $\{0.25, 0.5, 1, 2\}$. Thus $\mathbf{s}_i = [\mathbf{q}_i; \mathbf{u}_i]$ has $9 + 80 + 12 = 101$ dimensions.

### E.4 Strict-isolation boundary

The learner/item/state-disjoint model removes state multiplicity, disables the singleton correction, and zeros historical-strength-derived event channels and the parent-measurement relation. Test observations are predicted independently, and only Train/Validation states may provide residual labels.

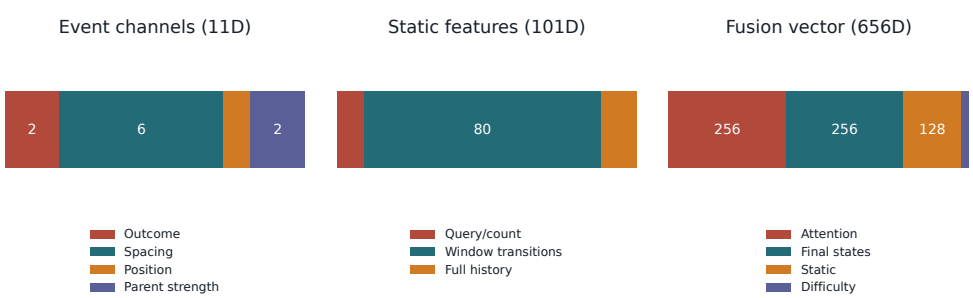


Figure 11: Dimensional composition of the event channels, static features, and 656-dimensional fused representation. Widths are normalized within each block; labels give exact dimensions.

# F Evaluation, Data, and Training Details

**Scope.** This appendix documents the shared prediction setup used by Section 5.2: data construction, label visibility, optimization, baseline adaptation, and fixed-split reporting. Held-out-user confirmation and source-only generalization details are separated into Appendices C and D.

## F.1 Information boundary

All four capability-control views share the same Train/Validation/Test splits and input-information boundary. Table 9 summarizes which prediction components each side receives. No view uses the current Test outcome

Table 9: Definition of the four capability-control views. "Exact" uses only protocol-visible Train/Validation labels; "residual" denotes the fixed relation-based postprocessor.

| View | Baseline path | OSR-HLR path | What the comparison isolates |
|---|---|---|---|
| label_free | Neural | Neural | Learned prediction without label retrieval |
| shared_exact | Neural + exact | Neural + exact | Neural prediction with identical exact lookup |
| equal_label_pool | Neural + exact | Neural + exact + residual | Primary system gain from residual transfer |
| equal_postprocessor | Neural + exact + residual | Neural + exact + residual | Encoder difference under matched postprocessing |

for training or checkpoint selection. On MaiMemo and Anki, however, these views are transductive. The split assigns aggregated observations rather than states, and released labels are fitted over all reviews of a state. When a Test observation shares its state with a Train or Validation observation (9.7% of Test observations on MaiMemo and 68.5% on Anki), exact lookup therefore returns a label that already reflects that Test observation's outcome. In the same views, OSR-HLR also reads the released per-event strength history and the state multiplicity. On MaiMemo, the last entry of that history equals the parent state's released label. These views are therefore reported as diagnostics only; headline claims use the source-only protocols (Appendix D).

## F.2 Capability-control results

In the stricter equal_postprocessor comparison, event/static inputs, difficulty embedding, output head, target, loss, optimization, split, exact-label pool, and residual-graph hyperparameters are matched; only encoder-internal fusion remains architecture-dependent.
Detailed view-wise metrics for EdNet-Recall are reported with the complete metric tables in Appendix H.

## F.3 Protocol, data, and implementation details

The capability-control views use deterministic outcome-independent partitions of the full released tables: MaiMemo and Anki split their aggregate rows 60/20/20. EdNet-Recall hashes delayed-response events 60/20/20 before the Train-only difficulty nuisance holdout, aggregation, and support filtering; claims requiring learner, item, state, exact-recurrence, or temporal boundaries use the corresponding dedicated track or diagnostic rather than the Row split.
Validation prediction uses Train labels only. The final frozen Test view uses Train and Validation labels, but is not used for checkpoint selection. Unlabeled Test covariates may define state membership, compatibility keys, observation counts, and aggregated neural predictions; residual values may come only from protocol-visible source states. In source-only evaluations, each Test observation is predicted independently and no Test state acts as a source. Perturbing current Test half-life or recall targets leaves every prediction unchanged. This test does not detect the shared-label effect described above, because it perturbs only the Test copy of a label.

**Sources and label construction.** MaiMemo is derived from the CC BY-NC 4.0 Student Memory Logs release associated with SSP-MMC and SSP-MMC-Plus (Ye et al., 2022; Su et al., 2023); Anki comes from the gated Anki Revlogs 10K release; and EdNet-Recall is derived from public EdNet-KT1 delayed repeated responses (Choi et al., 2020). We do not redistribute the raw logs; the artifact provides preprocessing code, source hashes, and split identifiers. Records are grouped by difficulty, outcome history, interval history, and elapsed-time

bin. Recall rates across bins are converted to a state-level half-life using Equation 2, weighted by the bin count. MaiMemo uses `total_cnt` for this weight and `group_cnt` for state support; Anki uses `group_cnt` as the bin weight; EdNet-Recall uses `total_cnt` as the cell weight and a split-local `group_cnt` as state support. All processed observations have finite positive labels.

**Partition construction.** The primary aggregate tables contain no learner, item, or absolute-time identifier. Strict MaiMemo evaluation therefore returns to the raw logs and uses salted hashes to create partitions with zero learner, item, and complete-state overlap. For Anki, a salted hash fixed 1,000 licensed user partitions before download; identity assignment hashes learner, collection-scoped card, and state independently, while the chronological split uses relative `day_offset` and within-day order. Labels are refitted within each split, and split-specific graph nodes prevent labels from being merged across partitions. EdNet-Recall hashes raw delayed-response events for the Row track and separately hashes learner, global question, and complete history state for the Learner, Item, and State tracks; Joint retains only rows whose three assignments agree. Its half-life labels are refitted within each split, and the split partition is part of the state identity.

**Implementation and optimization.** OSR-HLR uses $T_{\max} = 44$, $F_e = 11$, $F_s = 101$, width 128 for the event and static projections and each BiGRU direction, and width 16 for the difficulty embedding. Its $656 \to 256 \to 128 \to 1$ head gives 509,292 parameters, or 509,096 under the source-only boundary. The frozen state layer applies exact lookup followed by eight ordered residual relations: cross-difficulty, suffix-4, ancestor-2, suffix-8, difficulty slope, suffix-2, parent measurement, and sign-suffix-8. The training objective in Equation 8 uses $\lambda_h = 0.8$, SmoothL1 threshold $\beta = 0.5$, and $\varepsilon = 10^{-6}$. On MaiMemo, $w_i = \mathrm{clip}(c_i/\,\mathrm{median}_{\mathrm{Train}}\, c, 0.25, 4)^{1/2}$, $\tilde{w}_i = \mathrm{clip}(\log(1+c_i)/\,\mathrm{median}_{\mathrm{Train}} \log(1+c), 0.25, 4)$, and $\lambda_R = 10^{-3}$; Anki and EdNet-Recall use $w_i = \tilde{w}_i = 1$ and $\lambda_R = 0$. EdNet-Recall additionally disables historical-strength and state-multiplicity features, uses dropout 0.1, and selects checkpoints by raw Validation H-SMAPE. Optimization uses AdamW with learning rate $5 \times 10^{-4}$, weight decay $10^{-5}$, batch size 512, gradient clipping at 5, and ReduceLROnPlateau. Native-SRS runs use dropout 0.2. Models train for at most 100 epochs with patience 10; native fixed-split reruns use seeds 42–51 and EdNet-Recall uses seeds 42–44 on each track. All runs are CPU-only.

**Baseline details.** The cognitive/formulaic family contains Pimsleur and Leitner progression rules, ACT-R activation, HLR count regression, DASH spacing-weighted histories, and DHP/FSRS-style difficulty–stability transitions (Pimsleur, 1967; Leitner, 1972; Anderson et al., 1997; Settles & Meeder, 2016; Lindsey et al., 2014; Ye et al., 2022). The generic sequence family contains RNN, GRU, LSTM, Transformer, and DKT (Elman, 1990; Cho et al., 2014; Hochreiter & Schmidhuber, 1997; Vaswani et al., 2017; Piech et al., 2015). Memory-specific models add forgetting-aware attention or state dynamics: FIFAKT, GRU-HLR, THLR, ROKAN, and I-Mem (Ma et al., 2023; 2026; Su et al., 2023; Xiao & Wang, 2024; Shen et al., 2025a; 2026b). Recent neural models comprise EPsyINN, LefoKT, ReKT, LTKT, SPsyINN, KVFKT, GRKT, MIKT, MCKT, and DGMKT (Shen et al., 2026a; Bai et al., 2025; Shen et al., 2024; Xu et al., 2026; Sun et al., 2025; Guan et al., 2025; Cui et al., 2024; Sun et al., 2024; Shen et al., 2025b; Shao et al., 2025). All implementations retain the source model's central mechanism but use the common outcome/interval interface and half-life target. Inputs not consistently available across all three datasets, such as text, knowledge concepts, Q-matrices, or learner profiles, are omitted rather than reconstructed from Test information; the reported FSRS entry is therefore an FSRS-style task adaptation rather than a canonical scheduler reproduction.

## G Fixed-Scheduler Evaluation Protocols

**OSR-ACRI daily loop.** At day $t$, OSR-ACRI uses histories, due days $d_i$, and calibration state $c_{t-1}$ as follows:

1. Predict $\hat{h}_{i,t}$ and $\hat{p}_i(a_i)$ with OSR-HLR.
2. Form the due queue $\mathcal{Q}_t = \{i : d_i \le t\}$.
3. Compute $I_i(t)$ for every $i \in \mathcal{Q}_t$ using Equation 10.
4. Review the top $\min(K, |\mathcal{Q}_t|)$ items by $I_i(t)$.
5. Observe outcomes $y_i$ and append them to the corresponding histories.
6. Update $c_t$ and $g_t$ using Equation 9.
7. Re-predict $\hat{h}_{i,t}$ and set each reviewed item's next due day using Equation 9.

**Next-review decisions.** For target recall $q$, the shared threshold scheduler chooses

$$\hat{\Delta} = -\hat{h} \log_2 q. \tag{14}$$

The independently fitted reference half-life, rather than the model's own prediction, evaluates the decision: $p_i(q) = 2^{-\hat{\Delta}_i/h_i} = q^{\hat{h}_i/h_i}$. We report

$$E_q = \frac{1}{n}\sum_i |p_i(q) - q|, \qquad S_q = \frac{1}{n}\sum_i [q - p_i(q)]_+, \qquad U_q = \frac{1}{n}\sum_i [p_i(q) - q]_+,$$

for target-recall error, late-review shortfall, and early-review recall surplus, respectively; $E_q = S_q + U_q$. The surplus metric replaced a relative interval pressure whose target-recall factor cancels across the $q$ grid. The

Table 10: Next-review scheduling across target-recall levels: target error, recall shortfall, and early-review surplus for OSR-HLR against the matched baseline over six $q$ targets.

| Target $q$ | Relative reduction vs. matched baseline (%) | | |
|---|---|---|---|
| | Target error | Recall shortfall | Early-review surplus |
| 0.70 | **19.48%** | **22.64%** | **15.93%** |
| 0.75 | **19.54%** | **22.76%** | **15.84%** |
| 0.80 | **19.61%** | **22.89%** | **15.76%** |
| 0.85 | **19.67%** | **23.00%** | **15.68%** |
| **0.90** | **19.73%** | **23.12%** | **15.60%** |
| 0.95 | **19.80%** | **23.23%** | **15.54%** |

primary matched comparison uses seeds 43–51, state-cluster bootstrap intervals, and six targets from $q = 0.70$ to 0.95. Across all 27 baselines and six targets, OSR-HLR is lower in all 162 baseline–target comparisons. At $q = 0.90$, its reductions span **19.52–86.91%** for target error, **20.84–92.28%** for shortfall, and **15.60–78.58%** for surplus.

**Fixed-scheduler simulation.** Every method uses the same integer-day scheduler,

$$\Delta_t = \max\left\{1, \text{round}\left[-\hat{h}_t \log_2(0.90)\right]\right\}. \tag{15}$$

For held-out state $i$, source-only Test error is

$$e_i = \log \frac{\hat{h}_i}{h_i}, \tag{16}$$

and the candidate receives $\hat{h}_t = h_t \exp(e_t)$ at each simulated review. The *persistent* process preserves the observed Test-state error; *resampled* draws a paired observation index from the complete Validation error pool at every review. This transports measured error without repeatedly querying a model on out-of-distribution synthetic histories.

The evaluation uses 2,549 identity-disjoint Test states, four matched rollouts per state, a 365-day horizon, a 100-review cap, two source-fitted transition environments, and shared outcome uniforms and error indices. Crossing the two environments with the two error processes gives four scenarios. Each trajectory is compared with its matched perfect prediction reference using absolute review-count regret, target-recall MAE regret, retention-AUC regret, and time-below-target regret. Checkpoints first average rollouts within state and then scenarios; curves report the Test-state mean and sample SD.

The simulator design was selected on Validation. The initial Test comparison covered EPsyINN, LTKT, and DGMKT before the same evaluation was extended to all 27 baselines. OSR-HLR is lower on all four metrics in all 108 baseline–scenario comparisons. It is also lower at all 15 review/target checkpoints and all 14 nonzero AUC/time checkpoints. These counts summarize consistency within the frozen simulator; they are not independent samples or evidence of real-learner effects.

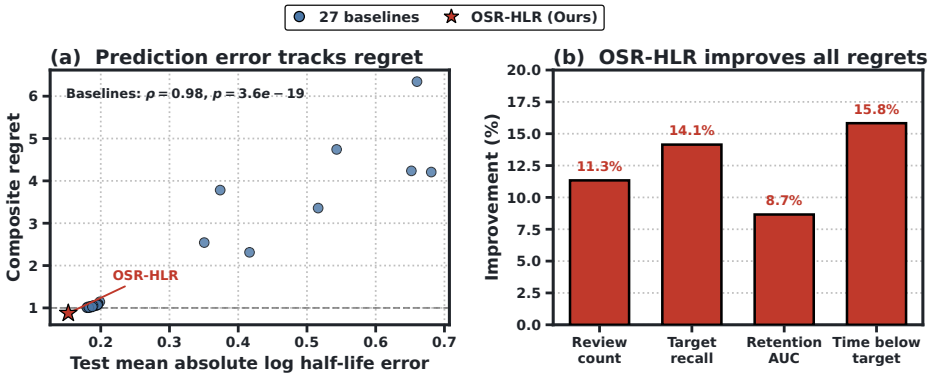


Figure 12: Held-out prediction error versus long-horizon scheduling regret across all 27 baselines and four scenarios; the two track closely ($\rho = 0.98$), linking more accurate half-life estimates to lower regret.

**Fixed-predictor OSR-ACRI evaluation.** The scheduler comparison fixes a three-seed OSR-HLR ensemble and varies only the scheduling rule under matched states, environments, capacity, and random streams. The Validation-selected setting uses target recall $q = 0.90$, calibration smoothing $\gamma = 0.90$, contraction strength $\kappa = 4$, priority scale $\beta = 12$, and daily capacity $K = 32$. Test evaluation uses 512 complete states, four rollouts, four source-domain environments, and a 365-day horizon. The protocol keeps total review workload matched to Threshold and reports utility across the reference schedulers and source-domain environments.

Table 11: Fixed-scheduler prediction comparison: regret relative to a perfect prediction reference under the shared threshold scheduler for all 27 baselines and OSR-HLR, aggregated over four prespecified scenarios.

| Method | Next review | | | 365-day regret vs. perfect prediction | | | |
|---|---|---|---|---|---|---|---|
| | Target err. | Shortfall | Early-review surplus | Review regret | Target regret | AUC regret | Time-below regret |
| *Cognitive / formulaic* | | | | | | | |
| Pimsleur | $5.75 \pm 0.01$ | $2.72 \pm 0.01$ | $3.03 \pm 0.00$ | $14.79 \pm 13.64$ | $7.56 \pm 5.96$ | $11.05 \pm 9.26$ | $25.63 \pm 12.92$ |
| Leitner | $5.38 \pm 0.00$ | $2.36 \pm 0.00$ | $3.02 \pm 0.00$ | $16.31 \pm 13.12$ | $6.76 \pm 3.99$ | $10.74 \pm 9.50$ | $25.43 \pm 12.04$ |
| ACT-R | $10.46 \pm 0.05$ | $9.33 \pm 0.07$ | $1.13 \pm 0.02$ | $11.60 \pm 8.35$ | $14.50 \pm 12.27$ | $19.28 \pm 12.58$ | $39.14 \pm 16.40$ |
| HLR | $7.40 \pm 0.07$ | $5.79 \pm 0.22$ | $1.61 \pm 0.16$ | $10.20 \pm 7.97$ | $10.31 \pm 9.40$ | $13.45 \pm 9.67$ | $31.47 \pm 14.08$ |
| DASH | $3.76 \pm 0.09$ | $1.87 \pm 0.06$ | $1.90 \pm 0.06$ | $8.06 \pm 8.96$ | $4.84 \pm 3.56$ | $4.98 \pm 5.87$ | $15.86 \pm 9.47$ |
| DHP | $5.73 \pm 0.01$ | $4.03 \pm 0.01$ | $1.69 \pm 0.00$ | $8.69 \pm 7.82$ | $7.77 \pm 5.59$ | $7.59 \pm 6.24$ | $25.32 \pm 11.07$ |
| FSRS | $4.15 \pm 0.09$ | $3.07 \pm 0.10$ | $1.08 \pm 0.03$ | $6.58 \pm 5.38$ | $5.99 \pm 6.29$ | $5.71 \pm 5.89$ | $19.00 \pm 10.69$ |
| *Generic sequence* | | | | | | | |
| RNN | $1.77 \pm 0.04$ | $0.92 \pm 0.05$ | $0.85 \pm 0.05$ | $3.66 \pm 3.56$ | $2.43 \pm 2.32$ | $1.90 \pm 2.50$ | $7.12 \pm 4.89$ |
| GRU | $1.77 \pm 0.02$ | $0.94 \pm 0.05$ | $0.83 \pm 0.05$ | $3.75 \pm 3.66$ | $2.44 \pm 2.33$ | $1.90 \pm 2.52$ | $7.28 \pm 4.92$ |
| LSTM | $1.78 \pm 0.01$ | $0.96 \pm 0.04$ | $0.82 \pm 0.03$ | $3.74 \pm 3.59$ | $2.44 \pm 2.33$ | $1.90 \pm 2.49$ | $7.33 \pm 4.91$ |
| Transformer | $1.95 \pm 0.15$ | $1.18 \pm 0.16$ | $0.78 \pm 0.06$ | $3.94 \pm 3.62$ | $2.66 \pm 2.53$ | $2.05 \pm 2.56$ | $8.70 \pm 5.67$ |
| DKT | $1.73 \pm 0.03$ | $0.91 \pm 0.04$ | $0.82 \pm 0.05$ | $3.65 \pm 3.66$ | $2.41 \pm 2.34$ | $1.89 \pm 2.56$ | $6.93 \pm 4.82$ |
| *Memory / forgetting* | | | | | | | |
| FIFAKT | $1.72 \pm 0.03$ | $0.94 \pm 0.06$ | $0.78 \pm 0.04$ | $3.61 \pm 3.55$ | $2.37 \pm 2.31$ | $1.86 \pm 2.47$ | $7.03 \pm 4.83$ |
| GRU-HLR | $1.76 \pm 0.02$ | $0.93 \pm 0.05$ | $0.83 \pm 0.03$ | $3.69 \pm 3.56$ | $2.44 \pm 2.31$ | $1.90 \pm 2.46$ | $7.27 \pm 4.86$ |
| THLR | $1.86 \pm 0.07$ | $0.99 \pm 0.09$ | $0.87 \pm 0.06$ | $3.77 \pm 3.65$ | $2.51 \pm 2.35$ | $1.96 \pm 2.57$ | $7.34 \pm 4.99$ |
| ROKAN | $4.86 \pm 0.28$ | $3.88 \pm 0.28$ | $0.98 \pm 0.08$ | $9.68 \pm 7.43$ | $6.40 \pm 7.81$ | $11.74 \pm 8.92$ | $23.71 \pm 11.36$ |
| I-Mem | $1.90 \pm 0.10$ | $1.04 \pm 0.16$ | $0.86 \pm 0.08$ | $3.90 \pm 3.74$ | $2.56 \pm 2.34$ | $2.00 \pm 2.59$ | $7.84 \pm 5.06$ |
| *Recent KT / memory* | | | | | | | |
| EPsyINN | $1.71 \pm 0.01$ | $0.94 \pm 0.02$ | $0.77 \pm 0.02$ | $3.64 \pm 3.52$ | $2.40 \pm 2.32$ | $1.89 \pm 2.51$ | $7.33 \pm 4.88$ |
| LefoKT | $1.87 \pm 0.03$ | $1.04 \pm 0.05$ | $0.83 \pm 0.05$ | $3.82 \pm 3.73$ | $2.57 \pm 2.36$ | $1.99 \pm 2.57$ | $7.85 \pm 5.15$ |
| ReKT | $1.80 \pm 0.03$ | $0.97 \pm 0.05$ | $0.83 \pm 0.03$ | $3.78 \pm 3.69$ | $2.53 \pm 2.35$ | $1.99 \pm 2.60$ | $7.49 \pm 5.02$ |
| LTKT | $1.70 \pm 0.03$ | $0.91 \pm 0.02$ | $0.79 \pm 0.04$ | $3.61 \pm 3.56$ | $2.37 \pm 2.29$ | $1.88 \pm 2.51$ | $7.05 \pm 4.77$ |
| SPsyINN | $1.71 \pm 0.03$ | $0.91 \pm 0.04$ | $0.80 \pm 0.02$ | $3.66 \pm 3.54$ | $2.40 \pm 2.32$ | $1.91 \pm 2.51$ | $7.21 \pm 4.81$ |
| KVFKT | $1.81 \pm 0.02$ | $0.99 \pm 0.03$ | $0.82 \pm 0.03$ | $3.77 \pm 3.61$ | $2.50 \pm 2.36$ | $1.97 \pm 2.53$ | $7.46 \pm 4.99$ |
| GRKT | $1.85 \pm 0.03$ | $1.01 \pm 0.04$ | $0.85 \pm 0.04$ | $3.82 \pm 3.61$ | $2.52 \pm 2.32$ | $1.97 \pm 2.51$ | $7.70 \pm 5.09$ |
| MIKT | $1.76 \pm 0.03$ | $0.96 \pm 0.02$ | $0.81 \pm 0.02$ | $3.73 \pm 3.55$ | $2.48 \pm 2.36$ | $1.95 \pm 2.52$ | $7.48 \pm 4.95$ |
| MCKT | $1.73 \pm 0.03$ | $0.92 \pm 0.04$ | $0.81 \pm 0.03$ | $3.62 \pm 3.52$ | $2.38 \pm 2.29$ | $1.88 \pm 2.49$ | $7.03 \pm 4.74$ |
| DGMKT | $1.76 \pm 0.02$ | $0.95 \pm 0.05$ | $0.82 \pm 0.04$ | $3.67 \pm 3.64$ | $2.43 \pm 2.32$ | $1.91 \pm 2.55$ | $7.14 \pm 4.91$ |
| *Ours* | | | | | | | |
| **OSR-HLR (Ours)** | $\mathbf{1.37 \pm 0.00}$ | $\mathbf{0.72 \pm 0.01}$ | $\mathbf{0.65 \pm 0.01}$ | $\mathbf{3.20 \pm 3.43}$ | $\mathbf{2.03 \pm 2.17}$ | $\mathbf{1.70 \pm 2.41}$ | $\mathbf{5.83 \pm 4.34}$ |
| **Improvement (%)** | **+19.5%** | **+20.8%** | **+15.6%** | **+11.3%** | **+14.1%** | **+8.7%** | **+15.8%** |

## H Metric Definitions and Reporting

**Statistical units.** Fixed-split values are means over ten independent training seeds on the same held-out observations, with sample SD across runs; the accompanying fixed-split uncertainty analysis uses the prespecified between-run comparison. Source-only generalization and held-out-user comparisons pair methods by seed and report paired intervals and exact sign tests within their prespecified comparison sets. One-step scheduling uses complete-state cluster bootstrap, while OSR-ACRI uses the complete initial state as the bootstrap unit. Repeated observations, checkpoints, baseline–scenario cells, and Monte Carlo trajectories are not treated as independent experimental replications.

**Primary estimand.** Unless stated otherwise, prediction results weight aggregated observations equally. This estimand differs from weighting learners equally or weighting by the number of underlying reviews. Those alternatives are reported as sensitivity analyses and cannot be inferred from observation-equal H-SMAPE alone. Let $y_i$ and $\hat{y}_i$ denote either the observed and predicted half-life or the observed and predicted recall probability. For $N$ Test examples, the complete tables use

$$\mathrm{MAE} = \frac{1}{N}\sum_i |\hat{y}_i - y_i|, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_i (\hat{y}_i - y_i)^2},$$

$$\mathrm{MAPE} = \frac{100}{N}\sum_i \frac{|\hat{y}_i - y_i|}{|y_i| + \epsilon}, \qquad \mathrm{SMAPE} = \frac{100}{N}\sum_i \frac{2|\hat{y}_i - y_i|}{|\hat{y}_i| + |y_i| + \epsilon},$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2 + \epsilon}.$$

MAE, RMSE, MAPE, and SMAPE are lower-is-better; $R^2$ is higher-is-better. H-MAE and H-RMSE inherit the day unit of half-life, whereas all P metrics compare values on the $[0, 1]$ probability scale. Negative $R^2$ values are valid and indicate that squared error exceeds the constant Test-mean predictor, which can occur for low-variance recall-probability targets.

The following compact tables report mean ± sample SD. Each table combines half-life and recall-probability metrics for one dataset and one capability-control view; MaiMemo and Anki use ten seeds, and EdNet-Recall averages the five source-only generalization protocols with three seeds per protocol. The MaiMemo and Anki tables are transductive (Appendix F). **Red bold**, blue underlined, and orange mark ranks 1–3.

# I Model-Agnostic Effect of Open-Space Retrieval

In the primary comparison (Table 2), every baseline receives the same source-only residual postprocessor as OSR-HLR. Table 25 isolates what this layer contributes on its own. For each baseline checkpoint, it compares Test H-SMAPE of the baseline's own neural prediction (w/o) with the same prediction after open-space retrieval (w/). Both columns use the source-only Row protocol: labels are refitted within each split, residual evidence comes only from Train and Validation states, and no checkpoint is retrained.

Retrieval lowers H-SMAPE for all 27 baselines on all three datasets, with median relative reductions of 8.28% on MaiMemo, 11.60% on Anki, and 4.94% on EdNet-Recall. The gain holds across all four model families, from formulaic rules to recent knowledge-tracing models, so open-space retrieval acts as a model-agnostic component rather than a correction tuned to one encoder. Because the primary comparison already gives every baseline this layer, the remaining margin of OSR-HLR in Table 2 reflects its relative-target base estimator; the core ablation in Section 5.2 quantifies the same effect within OSR-HLR.

Table 25: Model-agnostic effect of open-space retrieval under the source-only Row protocol. For each baseline, "w/o retrieval" is Test H-SMAPE of its own neural prediction and "w/ retrieval" adds the shared source-only residual postprocessor to the same checkpoint. Values are mean ± sample standard deviation over ten seeds (MaiMemo, Anki) or three seeds (EdNet-Recall). Red percentages give the relative H-SMAPE reduction from w/o to w/ retrieval.

| Method | MaiMemo | | Anki | | EdNet-Recall | |
|---|---|---|---|---|---|---|
| | w/o retrieval ↓ | w/ retrieval ↓ | w/o retrieval ↓ | w/ retrieval ↓ | w/o retrieval ↓ | w/ retrieval ↓ |
| *Cognitive / formulaic* | | | | | | |
| Pimsleur | 69.27±0.08 | 63.79±0.07 ↓**7.91%** | 116.64±0.01 | 100.02±0.01 ↓**14.24%** | 142.71±0.01 | 135.46±0.01 ↓**5.07%** |
| Leitner | 68.15±0.01 | 61.15±0.01 ↓**10.26%** | 110.28±0.01 | 94.93±0.00 ↓**13.92%** | 144.61±0.00 | 137.52±0.00 ↓**4.90%** |
| ACT-R | 56.37±0.06 | 51.70±0.04 ↓**8.28%** | 85.98±0.02 | 73.73±0.05 ↓**14.26%** | 85.44±0.01 | 77.52±0.02 ↓**9.26%** |
| HLR | 56.33±2.33 | 50.29±1.86 ↓**10.73%** | 84.78±2.65 | 70.44±2.03 ↓**16.92%** | 78.10±0.96 | 74.04±0.81 ↓**5.20%** |
| DASH | 47.63±1.25 | 42.29±1.12 ↓**11.22%** | 84.45±1.84 | 70.05±1.44 ↓**17.05%** | 80.37±0.90 | 75.67±0.69 ↓**5.85%** |
| DHP | 53.69±0.02 | 47.65±0.02 ↓**11.25%** | 85.27±0.05 | 72.06±0.07 ↓**15.49%** | 88.72±3.95 | 82.94±3.55 ↓**6.52%** |
| FSRS | 37.09±0.48 | 32.63±0.40 ↓**12.03%** | 62.17±1.70 | 52.25±1.33 ↓**15.97%** | 73.68±6.33 | 69.83±6.18 ↓**5.22%** |
| *Generic sequence* | | | | | | |
| RNN | 19.31±1.08 | 17.56±0.73 ↓**9.05%** | 45.84±1.48 | 40.77±1.04 ↓**11.05%** | 63.43±0.39 | 60.46±0.28 ↓**4.69%** |
| GRU | 18.23±0.32 | 16.81±0.21 ↓**7.75%** | 46.62±1.18 | 41.21±0.78 ↓**11.60%** | 74.29±1.64 | 72.05±1.27 ↓**3.02%** |
| LSTM | 18.54±0.75 | 17.03±0.48 ↓**8.16%** | 46.07±1.08 | 40.91±0.79 ↓**11.20%** | 72.02±8.44 | 69.30±8.10 ↓**3.78%** |
| Transformer | 19.68±1.00 | 17.55±0.70 ↓**10.84%** | 48.68±1.84 | 42.17±1.56 ↓**13.39%** | 76.57±4.54 | 70.59±3.79 ↓**7.82%** |
| DKT | 17.92±0.61 | 16.61±0.44 ↓**7.34%** | 43.81±1.32 | 40.12±0.85 ↓**8.43%** | 64.46±0.84 | 61.40±0.63 ↓**4.76%** |
| *Memory / forgetting* | | | | | | |
| FIFAKT | 18.50±0.66 | 16.80±0.47 ↓**9.17%** | 43.79±1.18 | 39.34±1.01 ↓**10.17%** | 65.65±0.66 | 62.17±0.62 ↓**5.29%** |
| GRU-HLR | 17.98±0.23 | 16.68±0.17 ↓**7.24%** | 46.16±1.19 | 40.88±0.82 ↓**11.45%** | 61.72±1.06 | 59.26±0.83 ↓**3.98%** |
| THLR | 19.23±0.49 | 17.51±0.32 ↓**8.97%** | 48.61±2.15 | 42.86±1.68 ↓**11.81%** | 66.49±0.86 | 62.72±0.54 ↓**5.67%** |
| ROKAN | 34.79±5.35 | 31.38±5.05 ↓**9.80%** | 54.62±2.35 | 47.48±1.76 ↓**13.07%** | 68.89±0.68 | 64.31±0.44 ↓**6.65%** |
| I-Mem | 21.42±1.52 | 18.94±1.00 ↓**11.55%** | 46.54±1.41 | 40.96±0.90 ↓**11.98%** | 63.59±1.15 | 60.73±0.71 ↓**4.49%** |
| *Recent KT / memory* | | | | | | |
| EPsyINN | 17.41±0.31 | 16.28±0.22 ↓**6.49%** | 44.92±1.08 | 40.33±0.69 ↓**10.20%** | 64.12±0.22 | 61.24±0.14 ↓**4.50%** |
| LefoKT | 19.47±0.43 | 17.66±0.36 ↓**9.30%** | 48.84±2.60 | 42.34±2.09 ↓**13.31%** | 72.02±0.84 | 66.91±0.51 ↓**7.10%** |
| ReKT | 18.48±0.30 | 17.09±0.24 ↓**7.52%** | 46.52±1.28 | 41.53±0.73 ↓**10.73%** | 64.64±0.46 | 61.64±0.37 ↓**4.64%** |
| LTKT | 17.44±0.61 | 16.26±0.51 ↓**6.73%** | 45.40±1.46 | 40.74±1.23 ↓**10.27%** | 65.54±0.89 | 62.18±0.74 ↓**5.13%** |
| SPsyINN | 17.94±0.50 | 16.68±0.42 ↓**7.04%** | 43.84±1.26 | 39.74±1.03 ↓**9.37%** | 62.50±0.91 | 59.99±0.74 ↓**4.02%** |
| KVFKT | 18.82±0.59 | 17.35±0.43 ↓**7.83%** | 49.98±1.31 | 43.63±0.71 ↓**12.70%** | 70.16±7.13 | 66.69±7.19 ↓**4.94%** |
| GRKT | 19.34±0.55 | 17.71±0.39 ↓**8.44%** | 47.58±2.48 | 42.07±1.50 ↓**11.59%** | 75.42±7.48 | 71.96±7.71 ↓**4.59%** |
| MIKT | 17.90±0.45 | 16.68±0.38 ↓**6.81%** | 46.72±1.31 | 41.70±0.80 ↓**10.74%** | 65.70±1.16 | 62.32±0.79 ↓**5.14%** |
| MCKT | 17.66±0.31 | 16.47±0.27 ↓**6.75%** | 45.16±1.64 | 40.35±1.10 ↓**10.67%** | 63.75±1.12 | 60.83±0.87 ↓**4.57%** |
| DGMKT | 18.10±0.36 | 16.83±0.23 ↓**7.02%** | 43.44±1.81 | 39.38±1.35 ↓**9.34%** | 63.89±1.07 | 60.99±0.93 ↓**4.54%** |
| **Summary** | Improved: 27/27 | median ↓**8.28%** | Improved: 27/27 | median ↓**11.60%** | Improved: 27/27 | median ↓**4.94%** |

# J Discussion, Limitations, and Future Work

**Scope of the evidence.** The prediction results cover the evaluated MaiMemo, Anki, and EdNet-Recall populations, label constructions, source-only generalization protocols, and held-out-user confirmation. Their different rating semantics, state recurrence, item identities, and data-collection policies explain why effect sizes vary across datasets and held-out estimands.

The two scheduling controls answer complementary questions. Fixed-scheduler perfect-prediction regret transports measured Test prediction error without repeatedly querying the model on synthetic histories. OSR-ACRI instead re-encodes every simulated outcome within the source-fitted transition family. OSR-ACRI is a local index rather than a globally optimal policy, and the component ablations indicate that delay-aware prioritization and queue age are its most robust scheduling mechanisms.

Table 12: equal_label_pool: MaiMemo half-life and recall-probability metrics. Values are mean ± sample SD; lower is better except $R^2$.

| Method | Half-life | | | | | Recall probability | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | H-SMAPE ↓ | H-MAE ↓ | H-RMSE ↓ | H-MAPE ↓ | H-$R^2$ ↑ | P-RMSE ↓ | P-MAE ↓ | P-SMAPE ↓ | P-MAPE ↓ | P-$R^2$ ↑ |
| *Cognitive / formulaic* | | | | | | | | | | |
| Pimsleur | 53.72±0.01 | 26.46±0.05 | 64.84±0.07 | 86.38±0.44 | 0.5277±0.0011 | 0.1492±0.0005 | 0.0956±0.0003 | 15.21±0.06 | 12.76±0.04 | −0.9280±0.0117 |
| Leitner | 48.18±0.00 | 18.94±0.02 | 44.51±0.01 | 59.30±0.10 | 0.7774±0.0001 | 0.1171±0.0000 | 0.0851±0.0000 | 10.93±0.00 | 11.99±0.00 | −0.1876±0.0004 |
| ACT-R | 49.17±0.01 | 30.50±0.19 | 108.59±1.24 | 393.37±3.45 | −0.3249±0.0303 | 0.0925±0.0000 | 0.0622±0.0000 | 7.88±0.00 | 8.55±0.00 | 0.2587±0.0007 |
| HLR | 48.26±0.02 | 22.61±0.12 | 56.12±0.33 | 180.71±1.31 | 0.6462±0.0042 | 0.0999±0.0005 | 0.0689±0.0002 | 10.04±0.06 | 9.40±0.03 | 0.1350±0.0086 |
| DASH | 36.37±0.03 | 17.09±0.13 | 49.38±0.65 | 41.79±0.24 | 0.7260±0.0074 | 0.0974±0.0003 | 0.0645±0.0001 | 9.65±0.03 | 8.85±0.02 | 0.1789±0.0046 |
| DHP | 26.47±0.01 | 13.26±0.01 | 38.46±0.13 | 28.40±0.23 | 0.8338±0.0011 | 0.0599±0.0001 | 0.0436±0.0001 | 5.69±0.01 | 5.76±0.01 | 0.6898±0.0008 |
| FSRS | 30.10±1.22 | 12.81±0.41 | 33.01±0.91 | 35.55±2.93 | 0.8775±0.0069 | 0.0699±0.0059 | 0.0497±0.0021 | 6.66±0.49 | 6.87±0.29 | 0.5739±0.0784 |
| *Generic sequence* | | | | | | | | | | |
| RNN | 14.71±0.42 | 7.98±0.31 | 25.57±1.37 | 14.57±0.44 | 0.9264±0.0080 | 0.0326±0.0009 | 0.0240±0.0007 | 3.13±0.10 | 3.13±0.09 | 0.9078±0.0053 |
| GRU | 14.27±0.23 | 7.64±0.13 | 24.52±0.62 | 14.13±0.13 | 0.9324±0.0034 | 0.0317±0.0004 | 0.0232±0.0004 | 3.02±0.05 | 3.03±0.05 | 0.9131±0.0021 |
| LSTM | 14.35±0.40 | 7.75±0.27 | 24.79±1.04 | 14.37±0.62 | 0.9309±0.0060 | 0.0319±0.0008 | 0.0233±0.0006 | 3.04±0.08 | 3.05±0.08 | 0.9117±0.0047 |
| Transformer | 15.95±0.59 | 8.72±0.50 | 27.17±1.96 | 16.10±0.78 | 0.9167±0.0126 | 0.0357±0.0013 | 0.0261±0.0010 | 3.44±0.13 | 3.45±0.14 | 0.8892±0.0079 |
| DKT | 14.32±0.19 | 7.68±0.14 | 24.48±0.86 | 14.38±0.24 | 0.9326±0.0048 | 0.0320±0.0005 | 0.0234±0.0004 | 3.06±0.06 | 3.08±0.06 | 0.9112±0.0030 |
| *Memory / forgetting* | | | | | | | | | | |
| FIFAKT | 14.73±0.39 | 8.06±0.28 | 25.70±1.13 | 14.80±0.44 | 0.9257±0.0066 | 0.0328±0.0009 | 0.0240±0.0007 | 3.14±0.11 | 3.15±0.10 | 0.9069±0.0054 |
| GRU-HLR | 14.61±0.52 | 7.84±0.39 | 24.83±1.32 | 14.61±0.61 | 0.9306±0.0076 | 0.0326±0.0012 | 0.0238±0.0008 | 3.11±0.11 | 3.12±0.11 | 0.9081±0.0072 |
| THLR | 15.99±0.53 | 8.51±0.29 | 25.95±0.56 | 15.97±0.67 | 0.9243±0.0033 | 0.0361±0.0017 | 0.0263±0.0011 | 3.46±0.17 | 3.46±0.17 | 0.8871±0.0107 |
| ROKAN | 34.64±2.25 | 19.95±1.98 | 57.85±4.79 | 60.33±40.32 | 0.6217±0.0646 | 0.1177±0.0378 | 0.0657±0.0118 | 10.04±3.04 | 8.72±1.52 | −0.3106±0.8825 |
| I-Mem | 14.69±0.21 | 7.87±0.15 | 25.12±0.59 | 14.67±0.44 | 0.9291±0.0033 | 0.0328±0.0006 | 0.0240±0.0004 | 3.14±0.06 | 3.16±0.06 | 0.9067±0.0032 |
| *Recent KT / memory* | | | | | | | | | | |
| EPsyINN | 14.22±0.25 | 7.65±0.19 | 24.63±0.83 | 14.02±0.19 | 0.9318±0.0046 | 0.0315±0.0005 | 0.0231±0.0004 | 3.00±0.05 | 3.01±0.05 | 0.9142±0.0026 |
| LefoKT | 16.22±0.52 | 8.62±0.31 | 26.41±1.27 | 16.44±0.45 | 0.9215±0.0076 | 0.0365±0.0012 | 0.0266±0.0009 | 3.50±0.12 | 3.52±0.11 | 0.8844±0.0080 |
| ReKT | 14.67±0.66 | 7.79±0.39 | 24.70±1.28 | 14.69±0.83 | 0.9313±0.0075 | 0.0327±0.0015 | 0.0239±0.0011 | 3.12±0.15 | 3.14±0.16 | 0.9073±0.0087 |
| LTKT | 14.48±0.21 | 7.70±0.18 | 24.68±0.84 | 14.38±0.24 | 0.9315±0.0047 | 0.0320±0.0005 | 0.0235±0.0004 | 3.06±0.05 | 3.07±0.05 | 0.9113±0.0027 |
| SPsyINN | 14.32±0.39 | 7.63±0.20 | 24.47±0.77 | 14.16±0.30 | 0.9327±0.0043 | 0.0318±0.0008 | 0.0234±0.0006 | 3.04±0.09 | 3.05±0.08 | 0.9121±0.0042 |
| KVFKT | 14.63±0.19 | 7.77±0.10 | 24.27±0.45 | 14.67±0.36 | 0.9338±0.0025 | 0.0325±0.0004 | 0.0238±0.0003 | 3.11±0.04 | 3.12±0.04 | 0.9086±0.0020 |
| GRKT | 15.28±0.69 | 8.11±0.27 | 25.03±0.93 | 15.26±0.62 | 0.9295±0.0053 | 0.0337±0.0012 | 0.0248±0.0011 | 3.24±0.14 | 3.24±0.13 | 0.9018±0.0070 |
| MIKT | 14.21±0.42 | 7.60±0.24 | 24.45±0.89 | 14.21±0.32 | 0.9327±0.0050 | 0.0317±0.0008 | 0.0231±0.0007 | 3.01±0.09 | 3.02±0.08 | 0.9131±0.0045 |
| MCKT | 14.46±0.25 | 7.77±0.23 | 24.63±1.08 | 14.55±0.40 | 0.9317±0.0061 | 0.0321±0.0007 | 0.0235±0.0005 | 3.07±0.07 | 3.09±0.07 | 0.9106±0.0037 |
| DGMKT | 14.42±0.37 | 7.67±0.20 | 23.97±0.48 | 14.53±0.49 | 0.9354±0.0026 | 0.0322±0.0009 | 0.0235±0.0006 | 3.07±0.09 | 3.09±0.10 | 0.9103±0.0050 |
| *Ours* | | | | | | | | | | |
| **OSR-HLR (Ours)** | **12.27±0.02** | **6.21±0.02** | **20.86±0.07** | **12.22±0.04** | **0.9511±0.0003** | **0.0273±0.0001** | **0.0198±0.0000** | **2.56±0.01** | **2.57±0.01** | **0.9353±0.0003** |
| **Improvement (%)** | **+13.67%** | **+18.26%** | **+13.00%** | **+12.86%** | **+1.68%** | **+13.20%** | **+14.42%** | **+14.75%** | **+14.54%** | **+2.32%** |

Table 13: equal_label_pool: Anki half-life and recall-probability metrics. Values are mean ± sample SD; lower is better except $R^2$.

| Method | Half-life | | | | | Recall probability | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | H-SMAPE ↓ | H-MAE ↓ | H-RMSE ↓ | H-MAPE ↓ | H-$R^2$ ↑ | P-RMSE ↓ | P-MAE ↓ | P-SMAPE ↓ | P-MAPE ↓ | P-$R^2$ ↑ |
| *Cognitive / formulaic* | | | | | | | | | | |
| Pimsleur | 39.42±0.02 | 26.09±0.43 | 432.16±9.92 | 764.10±13.06 | −3.2558±0.1957 | 0.1928±0.0004 | 0.1101±0.0002 | 29.33±0.06 | 19.18±0.02 | −0.1580±0.0045 |
| Leitner | 30.09±0.01 | 75.04±0.17 | 1369.06±2.13 | 1872.41±3.56 | −41.6901±0.1331 | 0.1731±0.0000 | 0.0987±0.0000 | 19.31±0.00 | 22.50±0.00 | 0.0669±0.0002 |
| ACT-R | 38.83±0.08 | 65.52±0.92 | 439.87±4.81 | 18339.93±276.41 | −3.4073±0.0963 | 0.1742±0.0001 | 0.1004±0.0001 | 19.32±0.02 | 21.76±0.02 | 0.0545±0.0015 |
| HLR | 38.17±0.14 | 18.67±0.36 | 186.24±5.46 | 3599.47±86.53 | 0.2094±0.0464 | 0.1722±0.0014 | 0.1012±0.0006 | 24.24±0.25 | 17.90±0.08 | 0.0759±0.0149 |
| DASH | 37.42±0.46 | 4.79±0.32 | 57.90±20.93 | 191.18±37.82 | 0.9147±0.0630 | 0.2419±0.0050 | 0.1326±0.0024 | 33.66±0.50 | 21.47±0.28 | −0.8237±0.0743 |
| DHP | 19.05±0.04 | 3.41±0.05 | 21.02±0.29 | 44.76±0.30 | 0.9899±0.0003 | 0.1269±0.0001 | 0.0748±0.0001 | 16.61±0.02 | 14.58±0.05 | 0.4985±0.0010 |
| FSRS | 19.90±0.35 | 3.32±0.15 | 24.37±1.58 | 49.65±4.76 | 0.9864±0.0018 | 0.1324±0.0007 | 0.0777±0.0004 | 17.27±0.09 | 16.46±0.08 | 0.4541±0.0054 |
| *Generic sequence* | | | | | | | | | | |
| RNN | 9.20±0.19 | 1.33±0.04 | 13.87±0.73 | 15.10±0.44 | 0.9956±0.0005 | 0.1175±0.0003 | 0.0620±0.0003 | 15.05±0.05 | 14.32±0.04 | 0.5700±0.0018 |
| GRU | 9.60±0.22 | 1.42±0.09 | 15.54±1.23 | 15.09±0.65 | 0.9945±0.0009 | 0.1180±0.0004 | 0.0627±0.0004 | 15.13±0.07 | 14.38±0.05 | 0.5666±0.0027 |
| LSTM | 9.65±0.31 | 1.45±0.08 | 15.65±0.86 | 15.13±0.73 | 0.9944±0.0006 | 0.1180±0.0003 | 0.0627±0.0004 | 15.13±0.06 | 14.41±0.08 | 0.5662±0.0024 |
| Transformer | 10.86±0.21 | 1.53±0.10 | 14.78±0.92 | 17.32±1.12 | 0.9950±0.0006 | 0.1201±0.0004 | 0.0648±0.0004 | 15.54±0.10 | 14.76±0.08 | 0.5504±0.0028 |
| DKT | 8.97±0.10 | 1.40±0.08 | 15.10±0.88 | 14.56±0.54 | 0.9948±0.0006 | 0.1173±0.0002 | 0.0616±0.0002 | 15.01±0.03 | 14.28±0.03 | 0.5715±0.0013 |
| *Memory / forgetting* | | | | | | | | | | |
| FIFAKT | 8.81±0.17 | 1.38±0.07 | 14.44±0.81 | 14.19±0.91 | 0.9952±0.0005 | 0.1170±0.0002 | 0.0612±0.0002 | 14.94±0.04 | 14.21±0.03 | 0.5735±0.0012 |
| GRU-HLR | 7.11±0.23 | 1.41±0.08 | 15.10±1.15 | 10.39±0.68 | 0.9948±0.0008 | 0.1046±0.0003 | 0.0543±0.0004 | 13.52±0.15 | 9.85±0.11 | 0.6593±0.0019 |
| THLR | 8.65±0.34 | 1.51±0.09 | 15.17±1.62 | 14.28±0.84 | 0.9947±0.0012 | 0.1071±0.0006 | 0.0570±0.0006 | 14.11±0.19 | 10.35±0.12 | 0.6425±0.0041 |
| ROKAN | 15.39±0.93 | 2.29±0.23 | 20.23±1.96 | 65.01±21.49 | 0.9906±0.0019 | 0.1306±0.0028 | 0.0722±0.0016 | 16.68±0.30 | 15.91±0.25 | 0.4686±0.0228 |
| I-Mem | 10.09±0.16 | 1.47±0.08 | 15.36±0.95 | 15.93±0.85 | 0.9946±0.0007 | 0.1185±0.0004 | 0.0635±0.0003 | 15.27±0.06 | 14.52±0.05 | 0.5626±0.0026 |
| *Recent KT / memory* | | | | | | | | | | |
| EPsyINN | 8.52±0.12 | 1.31±0.05 | 13.77±0.57 | 15.55±0.65 | 0.9957±0.0004 | 0.1166±0.0001 | 0.0605±0.0002 | 14.85±0.02 | 14.12±0.03 | 0.5768±0.0007 |
| LefoKT | 10.19±0.29 | 1.58±0.08 | 16.60±1.35 | 19.16±2.86 | 0.9937±0.0010 | 0.1188±0.0006 | 0.0635±0.0007 | 15.31±0.16 | 14.59±0.12 | 0.5602±0.0046 |
| ReKT | 9.47±0.11 | 1.60±0.10 | 16.73±1.39 | 20.37±1.64 | 0.9936±0.0011 | 0.1173±0.0002 | 0.0616±0.0002 | 14.98±0.03 | 14.34±0.04 | 0.5715±0.0014 |
| LTKT | 8.79±0.15 | 1.40±0.04 | 14.34±0.47 | 19.21±2.14 | 0.9953±0.0003 | 0.1168±0.0001 | 0.0607±0.0002 | 14.86±0.03 | 14.15±0.03 | 0.5752±0.0011 |
| SPsyINN | 8.40±0.08 | 1.33±0.06 | 14.42±0.83 | 15.75±0.87 | 0.9953±0.0005 | 0.1164±0.0001 | 0.0603±0.0001 | 14.81±0.01 | 14.09±0.02 | 0.5780±0.0004 |
| KVFKT | 9.36±0.11 | 1.52±0.05 | 15.30±0.49 | 20.46±0.74 | 0.9947±0.0003 | 0.1174±0.0001 | 0.0616±0.0002 | 14.98±0.03 | 14.28±0.04 | 0.5710±0.0010 |
| GRKT | 9.19±0.37 | 1.44±0.10 | 14.62±0.68 | 18.54±3.18 | 0.9951±0.0004 | 0.1170±0.0004 | 0.0612±0.0005 | 14.94±0.08 | 14.23±0.08 | 0.5732±0.0026 |
| MIKT | 8.74±0.20 | 1.44±0.08 | 15.44±0.77 | 17.74±1.38 | 0.9946±0.0005 | 0.1167±0.0002 | 0.0606±0.0003 | 14.85±0.03 | 14.13±0.04 | 0.5758±0.0012 |
| MCKT | 8.74±0.18 | 1.33±0.07 | 14.14±0.73 | 15.79±0.57 | 0.9954±0.0005 | 0.1167±0.0001 | 0.0608±0.0002 | 14.87±0.02 | 14.17±0.03 | 0.5760±0.0008 |
| DGMKT | 8.96±0.18 | 1.42±0.11 | 15.32±1.43 | 16.31±0.74 | 0.9946±0.0010 | 0.1168±0.0002 | 0.0610±0.0003 | 14.89±0.04 | 14.19±0.05 | 0.5748±0.0013 |
| *Ours* | | | | | | | | | | |
| **OSR-HLR (Ours)** | **4.63±0.01** | **0.70±0.01** | **7.62±0.19** | **4.86±0.02** | **0.9987±0.0001** | **0.1020±0.0000** | **0.0507±0.0000** | **12.77±0.00** | **9.11±0.01** | **0.6759±0.0002** |
| **Improvement (%)** | **+34.88%** | **+46.42%** | **+44.67%** | **+53.23%** | **+0.30%** | **+2.47%** | **+6.72%** | **+5.57%** | **+7.51%** | **+2.52%** |

Table 14: equal_label_pool: EdNet-Recall half-life and recall-probability metrics. Values are mean ± sample SD; lower is better except $R^2$.

| Method | Half-life | | | | | Recall probability | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | H-SMAPE ↓ | H-MAE ↓ | H-RMSE ↓ | H-MAPE ↓ | H-$R^2$ ↑ | P-RMSE ↓ | P-MAE ↓ | P-SMAPE ↓ | P-MAPE ↓ | P-$R^2$ ↑ |
| *Cognitive / formulaic* | | | | | | | | | | |
| Pimsleur | 158.68±14.73 | 287.66±68.42 | 467.51±143.97 | 99.39±3.67 | −0.5814±0.4390 | 0.4031±0.0771 | 0.3226±0.0900 | 84.81±30.00 | 58.01±11.04 | −3.8036±2.6724 |
| Leitner | 160.77±11.36 | 293.39±71.60 | 474.23±146.50 | 96.18±2.37 | −0.6161±0.4171 | 0.4049±0.0557 | 0.3232±0.0666 | 83.60±21.18 | 58.45±9.43 | −3.7181±1.9859 |
| ACT-R | 76.27±5.25 | 223.68±73.56 | 390.08±153.74 | 317.89±133.23 | 0.0439±0.1139 | 0.2925±0.0134 | 0.2263±0.0152 | 31.95±3.01 | 55.71±8.05 | −1.3634±0.4512 |
| HLR | 82.22±29.91 | 221.71±47.46 | 746.64±492.34 | 269.27±129.96 | −3.2473±3.7593 | 0.2987±0.0482 | 0.2289±0.0449 | 39.99±18.55 | 53.42±8.86 | −1.5822±1.2566 |
| DASH | 84.32±39.90 | 258.60±223.72 | 1003.04±2343.77 | 1674.04±5467.66 | −330.6153±1279.9400 | 0.3110±0.0802 | 0.2429±0.0828 | 45.16±31.46 | 54.40±10.90 | −1.9487±2.2642 |
| DHP | 81.50±19.27 | 226.34±48.08 | 373.54±104.30 | 350.47±154.57 | −0.3518±1.2117 | 0.2920±0.0122 | 0.2227±0.0143 | 35.13±5.12 | 53.79±7.66 | −1.3510±0.4128 |
| FSRS | 69.06±25.91 | 179.57±38.53 | 319.74±100.90 | 249.40±118.30 | 0.1512±0.5345 | 0.2813±0.0262 | 0.2166±0.0277 | 32.67±7.08 | 53.18±9.53 | −1.2201±0.6875 |
| *Generic sequence* | | | | | | | | | | |
| RNN | 55.47±10.44 | 157.72±49.66 | 281.01±117.26 | 168.29±55.01 | 0.4494±0.1887 | 0.2715±0.0147 | 0.2088±0.0165 | 30.04±3.01 | 47.79±6.82 | −1.0441±0.4503 |
| GRU | 59.09±13.64 | 161.69±52.70 | 283.70±116.39 | 174.58±61.46 | 0.4255±0.2177 | 0.2753±0.0181 | 0.2111±0.0179 | 31.03±3.51 | 50.39±6.88 | −1.0969±0.4513 |
| LSTM | 58.00±11.66 | 161.60±53.69 | 283.46±116.89 | 182.80±75.90 | 0.4399±0.1880 | 0.2737±0.0182 | 0.2102±0.0182 | 30.64±3.34 | 50.32±6.55 | −1.0745±0.4535 |
| Transformer | 61.29±10.52 | 188.35±65.76 | 326.88±126.87 | 269.71±125.47 | 0.2357±0.2808 | 0.2781±0.0163 | 0.2145±0.0176 | 30.52±3.00 | 53.20±8.49 | −1.1433±0.4717 |
| DKT | 55.36±11.38 | 154.80±49.48 | 275.06±110.38 | 180.06±70.24 | 0.4694±0.1750 | 0.2716±0.0156 | 0.2085±0.0171 | 29.92±3.09 | 52.01±8.55 | −1.0453±0.4561 |
| *Memory / forgetting* | | | | | | | | | | |
| FIFAKT | 53.85±11.30 | 158.02±53.49 | 286.22±117.71 | 177.25±114.96 | 0.4394±0.1662 | 0.2710±0.0174 | 0.2084±0.0195 | 29.87±3.04 | 51.82±8.94 | −1.0415±0.4934 |
| GRU-HLR | 54.63±10.02 | 151.48±44.24 | 270.96±107.46 | 166.84±71.61 | 0.4758±0.1900 | 0.2711±0.0167 | 0.2085±0.0186 | 29.97±3.15 | 50.00±7.41 | −1.0415±0.4770 |
| THLR | 56.01±9.64 | 158.57±52.55 | 278.54±103.71 | 219.18±109.00 | 0.4233±0.2499 | 0.2752±0.0172 | 0.2120±0.0198 | 30.16±2.99 | 52.70±8.95 | −1.1044±0.5028 |
| ROKAN | 64.77±11.07 | 202.67±70.12 | 350.59±121.94 | 281.68±115.09 | 0.0514±0.4649 | 0.2789±0.0156 | 0.2142±0.0163 | 31.12±3.38 | 53.04±8.88 | −1.1614±0.4994 |
| I-Mem | 60.42±13.78 | 167.94±46.07 | 311.25±108.30 | 194.81±93.27 | 0.2259±0.5144 | 0.2774±0.0277 | 0.2125±0.0255 | 31.46±6.53 | 52.25±8.78 | −1.1638±0.7237 |
| *Recent KT / memory* | | | | | | | | | | |
| EPsyINN | 55.34±10.19 | 153.09±45.67 | 269.50±104.85 | 192.11±62.25 | 0.4846±0.1855 | 0.2722±0.0158 | 0.2097±0.0183 | 29.81±3.06 | 52.06±8.26 | −1.0545±0.4588 |
| LefoKT | 61.57±13.87 | 180.50±40.92 | 299.50±68.69 | 305.45±167.08 | 0.0302±1.0288 | 0.2770±0.0198 | 0.2149±0.0236 | 30.36±3.44 | 53.30±9.19 | −1.1360±0.5472 |
| ReKT | 55.56±10.32 | 154.04±46.36 | 272.46±108.30 | 207.77±50.74 | 0.4760±0.1886 | 0.2724±0.0148 | 0.2099±0.0172 | 29.79±3.01 | 52.31±8.48 | −1.0563±0.4437 |
| LTKT | 56.32±11.18 | 157.82±40.97 | 279.10±92.77 | 232.92±111.54 | 0.3354±0.4643 | 0.2728±0.0166 | 0.2104±0.0191 | 29.86±3.11 | 52.41±8.82 | −1.0658±0.4790 |
| SPsyINN | 55.53±10.90 | 153.80±43.66 | 270.49±100.52 | 211.65±78.60 | 0.4506±0.2705 | 0.2727±0.0179 | 0.2105±0.0209 | 29.87±3.27 | 52.39±8.97 | −1.0690±0.5102 |
| KVFKT | 57.13±12.23 | 159.73±50.24 | 282.34±107.36 | 201.93±75.42 | 0.3909±0.3281 | 0.2735±0.0167 | 0.2107±0.0184 | 30.10±3.19 | 52.19±8.13 | −1.0723±0.4542 |
| GRKT | 58.64±14.11 | 162.14±55.14 | 283.94±121.67 | 224.05±66.00 | 0.4390±0.2035 | 0.2771±0.0200 | 0.2138±0.0215 | 30.77±3.78 | 52.88±8.82 | −1.1289±0.4936 |
| MIKT | 55.02±9.68 | 154.63±47.13 | 271.18±107.24 | 206.94±52.14 | 0.4866±0.1605 | 0.2719±0.0147 | 0.2094±0.0167 | 29.76±2.94 | 51.96±8.10 | −1.0464±0.4337 |
| MCKT | 55.27±10.18 | 155.43±44.28 | 277.82±103.52 | 219.12±89.83 | 0.4189±0.2867 | 0.2724±0.0165 | 0.2103±0.0195 | 29.79±3.08 | 52.32±8.85 | −1.0613±0.4824 |
| DGMKT | 55.42±9.58 | 153.24±47.28 | 274.96±105.43 | 207.90±64.90 | 0.4552±0.2109 | 0.2731±0.0175 | 0.2109±0.0213 | 29.93±3.12 | 52.47±8.87 | −1.0727±0.4994 |
| *Ours* | | | | | | | | | | |
| **OSR-HLR (Ours)** | **34.25±6.71** | **91.00±23.22** | **175.10±55.06** | **44.88±13.99** | **0.7629±0.1072** | **0.2569±0.0159** | **0.1950±0.0167** | **28.61±2.71** | **44.94±6.90** | **−0.8371±0.4485** |
| **Improvement (%)** | **+36.41%** | **+39.93%** | **+35.03%** | **+53.33%** | **+56.80%** | **+5.19%** | **+6.43%** | **+3.87%** | **+5.95%** | **+19.63%** |

Table 15: label_free: MaiMemo half-life and recall-probability metrics. Values are mean ± sample SD; lower is better except $R^2$.

| Method | Half-life | | | | | Recall probability | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | H-SMAPE ↓ | H-MAE ↓ | H-RMSE ↓ | H-MAPE ↓ | H-$R^2$ ↑ | P-RMSE ↓ | P-MAE ↓ | P-SMAPE ↓ | P-MAPE ↓ | P-$R^2$ ↑ |
| *Cognitive / formulaic* | | | | | | | | | | |
| Pimsleur | 63.69±0.01 | 37.72±0.04 | 92.84±0.05 | 93.04±0.44 | 0.0318±0.0011 | 0.1794±0.0005 | 0.1193±0.0003 | 18.69±0.08 | 15.44±0.04 | −1.7860±0.0170 |
| Leitner | 54.55±0.00 | 26.86±0.01 | 65.69±0.05 | 64.82±0.10 | 0.5152±0.0007 | 0.1245±0.0000 | 0.0939±0.0000 | 12.03±0.00 | 13.02±0.00 | −0.3419±0.0002 |
| ACT-R | 53.24±0.01 | 35.04±0.20 | 114.12±1.24 | 412.62±3.61 | −0.4632±0.0318 | 0.0958±0.0000 | 0.0658±0.0000 | 8.34±0.00 | 8.98±0.00 | 0.2048±0.0006 |
| HLR | 53.45±0.01 | 29.51±0.13 | 70.86±0.30 | 190.11±1.32 | 0.4360±0.0048 | 0.1026±0.0005 | 0.0735±0.0002 | 10.58±0.06 | 9.93±0.03 | 0.0877±0.0082 |
| DASH | 41.10±0.01 | 23.06±0.09 | 61.66±0.35 | 46.08±0.24 | 0.5728±0.0049 | 0.1012±0.0003 | 0.0696±0.0001 | 10.28±0.04 | 9.46±0.02 | 0.1136±0.0049 |
| DHP | 30.03±0.01 | 18.58±0.03 | 53.36±0.27 | 32.14±0.25 | 0.6802±0.0033 | 0.0624±0.0001 | 0.0466±0.0001 | 6.04±0.01 | 6.12±0.00 | 0.6632±0.0008 |
| FSRS | 32.77±1.25 | 16.09±0.53 | 40.97±1.84 | 38.39±2.97 | 0.8111±0.0173 | 0.0715±0.0057 | 0.0519±0.0020 | 6.92±0.49 | 7.15±0.29 | 0.5548±0.0781 |
| *Generic sequence* | | | | | | | | | | |
| RNN | 16.35±0.52 | 10.66±0.53 | 33.59±2.62 | 16.16±0.51 | 0.8726±0.0205 | 0.0336±0.0010 | 0.0249±0.0008 | 3.24±0.11 | 3.24±0.10 | 0.9022±0.0059 |
| GRU | 15.88±0.28 | 10.18±0.19 | 31.61±0.97 | 15.70±0.19 | 0.8876±0.0069 | 0.0327±0.0004 | 0.0241±0.0004 | 3.13±0.06 | 3.14±0.05 | 0.9074±0.0023 |
| LSTM | 15.97±0.54 | 10.33±0.49 | 32.06±2.14 | 15.95±0.75 | 0.8841±0.0165 | 0.0330±0.0010 | 0.0243±0.0007 | 3.16±0.09 | 3.17±0.10 | 0.9058±0.0059 |
| Transformer | 17.92±0.74 | 11.91±0.82 | 36.60±3.56 | 18.09±0.95 | 0.8482±0.0314 | 0.0370±0.0014 | 0.0273±0.0011 | 3.59±0.15 | 3.59±0.16 | 0.8815±0.0090 |
| DKT | 15.88±0.24 | 10.22±0.23 | 31.82±1.41 | 15.92±0.31 | 0.8860±0.0103 | 0.0330±0.0006 | 0.0243±0.0004 | 3.16±0.06 | 3.19±0.07 | 0.9059±0.0035 |
| *Memory / forgetting* | | | | | | | | | | |
| FIFAKT | 16.32±0.44 | 10.73±0.40 | 33.86±2.07 | 16.35±0.51 | 0.8708±0.0162 | 0.0337±0.0010 | 0.0249±0.0008 | 3.25±0.12 | 3.26±0.10 | 0.9015±0.0059 |
| GRU-HLR | 16.28±0.63 | 10.47±0.51 | 32.23±1.75 | 16.25±0.72 | 0.8830±0.0129 | 0.0336±0.0014 | 0.0248±0.0010 | 3.23±0.13 | 3.24±0.13 | 0.9021±0.0083 |
| THLR | 17.84±0.61 | 11.31±0.36 | 32.99±0.54 | 17.84±0.76 | 0.8777±0.0040 | 0.0372±0.0017 | 0.0274±0.0012 | 3.60±0.18 | 3.60±0.18 | 0.8799±0.0111 |
| ROKAN | 40.54±3.22 | 28.24±2.73 | 80.40±9.50 | 65.34±41.25 | 0.2646±0.1718 | 0.1357±0.0437 | 0.0768±0.0155 | 11.76±3.77 | 9.95±1.91 | −0.7448±1.1609 |
| I-Mem | 16.35±0.26 | 10.48±0.18 | 32.33±0.77 | 16.29±0.51 | 0.8825±0.0056 | 0.0339±0.0006 | 0.0250±0.0004 | 3.26±0.06 | 3.28±0.07 | 0.9006±0.0033 |
| *Recent KT / memory* | | | | | | | | | | |
| EPsyINN | 15.84±0.35 | 10.24±0.34 | 31.71±1.29 | 15.59±0.27 | 0.8869±0.0093 | 0.0325±0.0005 | 0.0240±0.0004 | 3.11±0.06 | 3.12±0.06 | 0.9086±0.0030 |
| LefoKT | 18.19±0.65 | 11.60±0.45 | 34.75±1.79 | 18.35±0.57 | 0.8640±0.0139 | 0.0382±0.0017 | 0.0279±0.0011 | 3.67±0.15 | 3.69±0.13 | 0.8735±0.0116 |
| ReKT | 16.45±0.81 | 10.57±0.59 | 32.34±1.64 | 16.44±1.00 | 0.8822±0.0122 | 0.0338±0.0016 | 0.0250±0.0012 | 3.25±0.16 | 3.27±0.18 | 0.9006±0.0100 |
| LTKT | 16.16±0.28 | 10.45±0.33 | 32.90±1.44 | 16.00±0.29 | 0.8782±0.0108 | 0.0334±0.0006 | 0.0246±0.0004 | 3.19±0.06 | 3.20±0.06 | 0.9031±0.0035 |
| SPsyINN | 15.99±0.52 | 10.27±0.36 | 31.60±1.15 | 15.77±0.44 | 0.8877±0.0083 | 0.0329±0.0008 | 0.0244±0.0007 | 3.16±0.10 | 3.17±0.09 | 0.9060±0.0047 |
| KVFKT | 16.36±0.24 | 10.49±0.12 | 31.64±0.68 | 16.38±0.42 | 0.8875±0.0049 | 0.0336±0.0004 | 0.0248±0.0003 | 3.23±0.04 | 3.24±0.05 | 0.9025±0.0023 |
| GRKT | 17.14±0.78 | 10.99±0.42 | 32.97±1.97 | 17.08±0.73 | 0.8775±0.0151 | 0.0350±0.0014 | 0.0260±0.0012 | 3.37±0.16 | 3.38±0.15 | 0.8939±0.0084 |
| MIKT | 15.91±0.57 | 10.28±0.42 | 31.48±1.17 | 15.88±0.47 | 0.8885±0.0084 | 0.0327±0.0010 | 0.0241±0.0008 | 3.12±0.11 | 3.14±0.10 | 0.9071±0.0056 |
| MCKT | 16.11±0.32 | 10.41±0.34 | 31.87±1.54 | 16.18±0.45 | 0.8856±0.0118 | 0.0331±0.0007 | 0.0245±0.0005 | 3.18±0.08 | 3.20±0.08 | 0.9052±0.0041 |
| DGMKT | 16.16±0.49 | 10.37±0.36 | 30.94±0.62 | 16.28±0.63 | 0.8924±0.0043 | 0.0332±0.0010 | 0.0245±0.0007 | 3.19±0.10 | 3.21±0.11 | 0.9042±0.0057 |
| *Ours* | | | | | | | | | | |
| **OSR-HLR (Ours)** | **12.95±0.06** | **7.14±0.06** | **21.90±0.16** | **12.83±0.08** | **0.9461±0.0008** | **0.0276±0.0001** | **0.0201±0.0001** | **2.60±0.01** | **2.61±0.01** | **0.9342±0.0005** |
| **Improvement (%)** | **+18.25%** | **+29.92%** | **+29.23%** | **+17.71%** | **+6.02%** | **+15.19%** | **+16.43%** | **+16.41%** | **+16.32%** | **+2.83%** |

Table 16: label_free: Anki half-life and recall-probability metrics. Values are mean ± sample SD; lower is better except $R^2$.

| Method | Half-life | | | | | Recall probability | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | H-SMAPE ↓ | H-MAE ↓ | H-RMSE ↓ | H-MAPE ↓ | H-$R^2$ ↑ | P-RMSE ↓ | P-MAE ↓ | P-SMAPE ↓ | P-MAPE ↓ | P-$R^2$ ↑ |
| *Cognitive / formulaic* | | | | | | | | | | |
| Pimsleur | 98.44±0.03 | 91.46±0.61 | 477.84±9.21 | 932.18±14.35 | −4.2021±0.2007 | 0.2793±0.0006 | 0.1881±0.0005 | 44.98±0.14 | 29.21±0.06 | −1.4305±0.0105 |
| Leitner | 99.18±0.00 | 158.13±0.25 | 1403.10±2.21 | 2044.88±3.87 | −43.8391±0.1413 | 0.3327±0.0001 | 0.2274±0.0001 | 42.70±0.02 | 37.86±0.01 | −2.4474±0.0016 |
| ACT-R | 89.87±0.03 | 126.00±0.97 | 478.68±4.60 | 1919.11±283.40 | −4.2192±0.1002 | 0.2820±0.0005 | 0.1714±0.0004 | 33.09±0.16 | 31.21±0.04 | −1.4769±0.0093 |
| HLR | 84.60±0.05 | 66.30±0.93 | 250.29±4.63 | 3840.56±85.08 | −0.4272±0.0528 | 0.2501±0.0007 | 0.1570±0.0004 | 34.55±0.13 | 25.00±0.05 | −0.9479±0.0107 |
| DASH | 86.56±0.14 | 50.42±0.94 | 176.37±8.01 | 296.06±36.75 | 0.2902±0.0655 | 0.3164±0.0037 | 0.2003±0.0021 | 46.35±0.41 | 29.78±0.24 | −2.1179±0.0723 |
| DHP | 71.17±0.01 | 59.08±0.07 | 179.78±0.19 | 92.41±0.55 | 0.2639±0.0015 | 0.2660±0.0002 | 0.1587±0.0002 | 35.20±0.05 | 25.12±0.05 | −1.2046±0.0029 |
| FSRS | 49.23±0.48 | 37.75±0.66 | 148.42±0.80 | 92.80±4.96 | 0.4983±0.0055 | 0.1507±0.0020 | 0.0958±0.0012 | 19.80±0.29 | 18.79±0.17 | 0.2925±0.0185 |
| *Generic sequence* | | | | | | | | | | |
| RNN | 19.35±0.43 | 19.03±0.45 | 137.88±0.67 | 25.47±0.52 | 0.5670±0.0042 | 0.1233±0.0016 | 0.0661±0.0010 | 15.75±0.22 | 14.95±0.13 | 0.5265±0.0124 |
| GRU | 20.28±0.54 | 18.89±0.39 | 137.88±0.47 | 25.88±1.01 | 0.5670±0.0029 | 0.1245±0.0011 | 0.0674±0.0008 | 15.94±0.20 | 15.09±0.11 | 0.5171±0.0087 |
| LSTM | 20.16±0.74 | 18.79±0.41 | 137.61±0.43 | 25.79±1.09 | 0.5687±0.0027 | 0.1247±0.0017 | 0.0674±0.0011 | 15.96±0.23 | 15.12±0.17 | 0.5156±0.0129 |
| Transformer | 24.40±0.65 | 22.34±0.85 | 140.06±1.25 | 31.70±1.64 | 0.5532±0.0080 | 0.1266±0.0023 | 0.0702±0.0016 | 16.59±0.42 | 15.61±0.24 | 0.5006±0.0183 |
| DKT | 18.85±0.33 | 18.63±0.29 | 137.72±0.23 | 24.64±0.61 | 0.5680±0.0015 | 0.1229±0.0011 | 0.0657±0.0007 | 15.69±0.15 | 14.91±0.09 | 0.5298±0.0086 |
| *Memory / forgetting* | | | | | | | | | | |
| FIFAKT | 18.84±0.52 | 19.06±0.51 | 138.14±0.58 | 24.30±1.02 | 0.5654±0.0036 | 0.1240±0.0014 | 0.0660±0.0009 | 15.85±0.18 | 14.96±0.13 | 0.5208±0.0110 |
| GRU-HLR | 17.18±0.44 | 18.97±0.30 | 137.66±0.33 | 20.48±0.92 | 0.5684±0.0020 | 0.1100±0.0011 | 0.0579±0.0006 | 14.09±0.17 | 10.31±0.10 | 0.6228±0.0077 |
| THLR | 20.35±0.81 | 20.42±0.98 | 138.68±0.85 | 26.30±1.35 | 0.5620±0.0053 | 0.1137±0.0028 | 0.0614±0.0017 | 14.87±0.37 | 10.93±0.25 | 0.5973±0.0197 |
| ROKAN | 38.54±2.86 | 28.84±2.05 | 145.77±1.74 | 138.70±35.18 | 0.5160±0.0115 | 0.1480±0.0050 | 0.0858±0.0033 | 18.78±0.65 | 17.77±0.50 | 0.3173±0.0460 |
| I-Mem | 21.39±0.53 | 19.53±0.58 | 137.70±0.60 | 27.48±1.15 | 0.5681±0.0037 | 0.1248±0.0020 | 0.0683±0.0011 | 16.05±0.32 | 15.24±0.17 | 0.5147±0.0152 |
| *Recent KT / memory* | | | | | | | | | | |
| EPsyINN | 18.15±0.25 | 18.58±0.49 | 137.44±0.50 | 25.37±0.67 | 0.5698±0.0031 | 0.1225±0.0010 | 0.0646±0.0005 | 15.57±0.13 | 14.77±0.07 | 0.5322±0.0078 |
| LefoKT | 22.86±0.91 | 20.82±0.78 | 138.91±0.85 | 32.70±2.60 | 0.5605±0.0054 | 0.1235±0.0039 | 0.0675±0.0027 | 15.93±0.67 | 15.21±0.43 | 0.5246±0.0301 |
| ReKT | 19.88±0.29 | 19.03±0.30 | 138.27±0.45 | 31.18±1.78 | 0.5645±0.0028 | 0.1232±0.0012 | 0.0660±0.0007 | 15.65±0.16 | 15.00±0.11 | 0.5270±0.0089 |
| LTKT | 18.87±0.47 | 18.78±0.55 | 137.99±0.83 | 29.52±2.42 | 0.5663±0.0052 | 0.1229±0.0011 | 0.0651±0.0006 | 15.58±0.17 | 14.81±0.10 | 0.5298±0.0083 |
| SPsyINN | 17.99±0.15 | 18.23±0.25 | 137.31±0.54 | 25.60±0.91 | 0.5706±0.0034 | 0.1220±0.0009 | 0.0641±0.0005 | 15.44±0.10 | 14.68±0.06 | 0.5365±0.0067 |
| KVFKT | 19.59±0.38 | 18.62±0.23 | 137.85±0.28 | 31.05±0.96 | 0.5672±0.0018 | 0.1238±0.0011 | 0.0660±0.0007 | 15.76±0.17 | 14.97±0.09 | 0.5226±0.0086 |
| GRKT | 19.66±0.84 | 18.76±0.37 | 137.93±0.58 | 29.39±3.65 | 0.5667±0.0037 | 0.1240±0.0012 | 0.0660±0.0010 | 15.80±0.18 | 14.96±0.14 | 0.5207±0.0091 |
| MIKT | 18.56±0.50 | 18.18±0.25 | 137.39±0.24 | 27.79±1.70 | 0.5701±0.0015 | 0.1238±0.0009 | 0.0652±0.0006 | 15.71±0.11 | 14.86±0.08 | 0.5229±0.0067 |
| MCKT | 18.63±0.26 | 18.53±0.40 | 137.64±0.38 | 25.99±0.68 | 0.5685±0.0024 | 0.1222±0.0008 | 0.0647±0.0004 | 15.52±0.11 | 14.78±0.06 | 0.5348±0.0058 |
| DGMKT | 18.74±0.31 | 18.47±0.36 | 137.71±0.36 | 26.44±0.92 | 0.5681±0.0023 | 0.1224±0.0014 | 0.0649±0.0007 | 15.53±0.23 | 14.80±0.12 | 0.5336±0.0108 |
| *Ours* | | | | | | | | | | |
| **OSR-HLR (Ours)** | **9.82±0.24** | **10.75±0.16** | **130.53±0.53** | **10.15±0.30** | **0.6119±0.0031** | **0.1026±0.0008** | **0.0516±0.0004** | **12.88±0.10** | **9.28±0.06** | **0.6718±0.0051** |
| **Improvement (%)** | **+42.85%** | **+40.88%** | **+4.94%** | **+50.41%** | **+7.25%** | **+6.72%** | **+10.98%** | **+8.64%** | **+9.94%** | **+7.87%** |

Table 17: label_free: EdNet-Recall half-life and recall-probability metrics. Values are mean ± sample SD; lower is better except $R^2$.

| Method | Half-life | | | | | Recall probability | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | H-SMAPE ↓ | H-MAE ↓ | H-RMSE ↓ | H-MAPE ↓ | H-$R^2$ ↑ | P-RMSE ↓ | P-MAE ↓ | P-SMAPE ↓ | P-MAPE ↓ | P-$R^2$ ↑ |
| *Cognitive / formulaic* | | | | | | | | | | |
| Pimsleur | 158.68±14.73 | 287.66±68.42 | 467.51±143.97 | 99.39±3.67 | −0.5814±0.4390 | 0.4031±0.0771 | 0.3226±0.0900 | 84.81±30.00 | 58.01±11.04 | −3.8036±2.6724 |
| Leitner | 160.77±11.36 | 293.39±71.60 | 474.23±146.50 | 96.18±2.37 | −0.6161±0.4171 | 0.4049±0.0557 | 0.3232±0.0666 | 83.60±21.18 | 58.45±9.43 | −3.7181±1.9859 |
| ACT-R | 76.27±5.25 | 223.68±73.56 | 390.08±153.74 | 317.89±133.23 | 0.0439±0.1139 | 0.2925±0.0134 | 0.2263±0.0152 | 31.95±3.01 | 55.71±8.05 | −1.3634±0.4512 |
| HLR | 82.22±29.91 | 221.71±47.46 | 746.64±492.34 | 269.27±129.96 | −3.2473±3.7593 | 0.2987±0.0482 | 0.2289±0.0449 | 39.99±18.55 | 53.42±8.86 | −1.5822±1.2566 |
| DASH | 84.32±39.90 | 258.57±223.72 | 1003.04±2343.77 | 1674.04±5467.66 | −330.6153±1279.9400 | 0.3110±0.0802 | 0.2450±0.0828 | 45.16±31.46 | 54.40±10.90 | −1.0487±2.2642 |
| DHP | 81.50±19.27 | 226.34±48.08 | 373.54±104.30 | 350.47±154.57 | −0.3518±1.2117 | 0.2920±0.0122 | 0.2227±0.0143 | 35.13±5.12 | 53.79±7.66 | −1.3510±0.4128 |
| FSRS | 69.06±25.91 | 179.57±38.53 | 319.74±100.90 | 249.40±118.30 | 0.1512±0.5345 | 0.2813±0.0262 | 0.2166±0.0277 | 32.67±7.08 | 53.18±9.53 | −1.2201±0.6875 |
| *Generic sequence* | | | | | | | | | | |
| RNN | 55.47±10.44 | 157.72±49.66 | 281.01±117.26 | 168.29±55.01 | 0.4494±0.1887 | 0.2715±0.0147 | 0.2088±0.0165 | 30.04±3.01 | 47.79±6.82 | −1.0441±0.4503 |
| GRU | 59.09±13.64 | 161.69±52.70 | 283.70±116.39 | 174.58±61.46 | 0.4255±0.2177 | 0.2753±0.0181 | 0.2111±0.0179 | 31.03±3.51 | 50.39±6.88 | −1.0969±0.4513 |
| LSTM | 58.00±11.66 | 161.60±53.69 | 283.46±116.89 | 182.80±75.90 | 0.4399±0.1880 | 0.2737±0.0182 | 0.2102±0.0182 | 30.64±3.34 | 50.32±6.55 | −1.0745±0.4535 |
| Transformer | 61.29±10.52 | 188.35±65.76 | 326.88±126.87 | 269.71±125.47 | 0.2357±0.2808 | 0.2781±0.0163 | 0.2145±0.0176 | 30.52±3.00 | 53.20±8.49 | −1.1433±0.4717 |
| DKT | 55.36±11.38 | 154.80±49.48 | 275.06±110.38 | 180.06±70.24 | 0.4694±0.1750 | 0.2716±0.0156 | 0.2085±0.0171 | 29.92±3.09 | 52.01±8.55 | −1.0453±0.4561 |
| *Memory / forgetting* | | | | | | | | | | |
| FIFAKT | 53.85±11.30 | 158.02±53.49 | 286.22±117.71 | 177.25±114.96 | 0.4394±0.1662 | 0.2710±0.0174 | 0.2084±0.0195 | 29.87±3.04 | 51.82±8.94 | −1.0415±0.4934 |
| GRU-HLR | 54.63±10.02 | 151.48±44.24 | 270.96±107.46 | 166.84±71.61 | 0.4758±0.1900 | 0.2711±0.0167 | 0.2085±0.0186 | 29.97±3.15 | 50.00±7.41 | −1.0415±0.4770 |
| THLR | 56.01±9.64 | 158.57±52.55 | 278.54±103.71 | 219.18±109.00 | 0.4233±0.2499 | 0.2752±0.0172 | 0.2120±0.0198 | 30.16±2.99 | 52.70±8.95 | −1.1044±0.5028 |
| ROKAN | 64.77±11.07 | 202.67±70.12 | 350.59±121.94 | 281.68±115.09 | 0.0514±0.4649 | 0.2789±0.0166 | 0.2142±0.0163 | 31.12±3.38 | 53.04±8.88 | −1.1614±0.4994 |
| I-Mem | 60.42±13.78 | 167.94±46.07 | 311.25±108.30 | 194.81±93.27 | 0.2259±0.5144 | 0.2774±0.0277 | 0.2125±0.0255 | 31.46±6.53 | 52.25±8.78 | −1.1638±0.7237 |
| *Recent KT / memory* | | | | | | | | | | |
| EPsyINN | 55.34±10.19 | 153.09±45.67 | 269.50±104.85 | 192.11±62.25 | 0.4846±0.1855 | 0.2722±0.0158 | 0.2097±0.0183 | 29.81±3.06 | 52.06±8.26 | −1.0545±0.4588 |
| LefoKT | 61.57±13.87 | 180.50±40.92 | 299.50±68.69 | 305.45±167.08 | 0.0302±1.0288 | 0.2770±0.0198 | 0.2149±0.0236 | 30.36±3.44 | 53.30±9.19 | −1.1360±0.5472 |
| ReKT | 55.56±10.32 | 154.04±46.36 | 272.46±108.30 | 207.77±50.74 | 0.4760±0.1886 | 0.2724±0.0148 | 0.2099±0.0172 | 29.79±3.01 | 52.31±8.48 | −1.0563±0.4437 |
| LTKT | 56.32±11.18 | 157.82±40.97 | 279.10±92.77 | 232.92±111.54 | 0.3354±0.4643 | 0.2728±0.0166 | 0.2104±0.0191 | 29.86±3.11 | 52.41±8.82 | −1.0658±0.4790 |
| SPsyINN | 55.53±10.90 | 153.80±43.66 | 270.49±100.52 | 211.65±78.60 | 0.4506±0.2705 | 0.2727±0.0179 | 0.2105±0.0209 | 29.87±3.27 | 52.39±8.97 | −1.0690±0.5102 |
| KVFKT | 57.13±12.23 | 159.73±50.24 | 282.34±107.36 | 201.93±75.42 | 0.3909±0.3281 | 0.2735±0.0167 | 0.2107±0.0184 | 30.10±3.19 | 52.19±8.13 | −1.0723±0.4542 |
| GRKT | 58.64±14.11 | 162.14±55.14 | 283.94±121.67 | 224.05±66.00 | 0.4390±0.2035 | 0.2771±0.0200 | 0.2138±0.0215 | 30.77±3.78 | 52.88±8.82 | −1.1289±0.4936 |
| MIKT | 55.02±9.68 | 154.63±47.13 | 271.18±107.24 | 206.94±52.14 | 0.4866±0.1605 | 0.2719±0.0147 | 0.2094±0.0167 | 29.76±2.94 | 51.96±8.10 | −1.0464±0.4337 |
| MCKT | 55.27±10.18 | 155.43±44.28 | 277.82±103.52 | 219.12±89.83 | 0.4189±0.2867 | 0.2724±0.0165 | 0.2103±0.0195 | 29.79±3.08 | 52.32±8.85 | −1.0613±0.4824 |
| DGMKT | 55.42±9.58 | 153.24±47.28 | 274.96±105.43 | 207.90±64.90 | 0.4552±0.2109 | 0.2731±0.0175 | 0.2109±0.0213 | 29.93±3.12 | 52.47±8.87 | −1.0727±0.4994 |
| *Ours* | | | | | | | | | | |
| **OSR-HLR (Ours)** | **35.16±6.61** | **93.75±23.49** | **178.99±54.45** | **47.38±14.36** | **0.7536±0.1059** | **0.2573±0.0159** | **0.1953±0.0168** | **28.65±2.74** | **45.04±6.91** | **−0.8418±0.4498** |
| **Improvement (%)** | **+34.70%** | **+38.11%** | **+33.59%** | **+50.73%** | **+54.89%** | **+5.06%** | **+6.25%** | **+3.76%** | **+5.75%** | **+19.17%** |

Table 18: shared_exact: MaiMemo half-life and recall-probability metrics. Values are mean ± sample SD; lower is better except $R^2$.

| Method | Half-life | | | | | Recall probability | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | H-SMAPE ↓ | H-MAE ↓ | H-RMSE ↓ | H-MAPE ↓ | H-$R^2$ ↑ | P-RMSE ↓ | P-MAE ↓ | P-SMAPE ↓ | P-MAPE ↓ | P-$R^2$ ↑ |
| *Cognitive / formulaic* | | | | | | | | | | |
| Pimsleur | 53.72±0.01 | 26.46±0.05 | 64.84±0.07 | 86.38±0.44 | 0.5277±0.0011 | 0.1492±0.0005 | 0.0956±0.0003 | 15.21±0.06 | 12.76±0.04 | −0.9280±0.0117 |
| Leitner | 48.18±0.00 | 18.94±0.02 | 44.51±0.01 | 59.30±0.10 | 0.7774±0.0001 | 0.1171±0.0000 | 0.0851±0.0000 | 10.93±0.00 | 11.99±0.00 | −0.1876±0.0004 |
| ACT-R | 49.17±0.01 | 30.50±0.19 | 108.59±1.24 | 393.37±3.45 | −0.3249±0.0303 | 0.0925±0.0000 | 0.0622±0.0000 | 7.88±0.00 | 8.55±0.00 | 0.2587±0.0007 |
| HLR | 48.26±0.02 | 22.61±0.12 | 56.12±0.33 | 180.71±1.31 | 0.6462±0.0042 | 0.0999±0.0005 | 0.0689±0.0002 | 10.04±0.06 | 9.40±0.03 | 0.1350±0.0086 |
| DASH | 36.37±0.03 | 17.09±0.13 | 49.38±0.65 | 41.79±0.24 | 0.7260±0.0074 | 0.0974±0.0003 | 0.0645±0.0001 | 9.65±0.03 | 8.85±0.02 | 0.1789±0.0046 |
| DHP | 26.47±0.01 | 13.26±0.01 | 38.46±0.13 | 28.40±0.23 | 0.8338±0.0011 | 0.0599±0.0001 | 0.0436±0.0001 | 5.69±0.01 | 5.76±0.01 | 0.6898±0.0008 |
| FSRS | 30.10±1.22 | 12.81±0.41 | 33.01±0.91 | 35.55±2.93 | 0.8775±0.0069 | 0.0699±0.0059 | 0.0497±0.0021 | 6.66±0.49 | 6.87±0.29 | 0.5739±0.0784 |
| *Generic sequence* | | | | | | | | | | |
| RNN | 14.71±0.42 | 7.98±0.31 | 25.57±1.37 | 14.57±0.44 | 0.9264±0.0080 | 0.0326±0.0009 | 0.0240±0.0007 | 3.13±0.10 | 3.13±0.09 | 0.9078±0.0053 |
| GRU | 14.27±0.23 | 7.64±0.13 | 24.52±0.62 | 14.13±0.13 | 0.9324±0.0034 | 0.0317±0.0004 | 0.0232±0.0004 | 3.02±0.05 | 3.03±0.05 | 0.9131±0.0021 |
| LSTM | 14.35±0.40 | 7.75±0.27 | 24.79±1.04 | 14.37±0.62 | 0.9309±0.0060 | 0.0319±0.0008 | 0.0233±0.0006 | 3.04±0.08 | 3.05±0.08 | 0.9117±0.0047 |
| Transformer | 15.95±0.59 | 8.72±0.50 | 27.17±1.96 | 16.10±0.78 | 0.9167±0.0126 | 0.0357±0.0013 | 0.0261±0.0010 | 3.44±0.13 | 3.45±0.14 | 0.8892±0.0079 |
| DKT | 14.32±0.19 | 7.68±0.14 | 24.48±0.86 | 14.38±0.24 | 0.9326±0.0048 | 0.0320±0.0005 | 0.0234±0.0004 | 3.06±0.06 | 3.08±0.06 | 0.9112±0.0030 |
| *Memory / forgetting* | | | | | | | | | | |
| FIFAKT | 14.73±0.39 | 8.06±0.28 | 25.70±1.13 | 14.80±0.44 | 0.9257±0.0066 | 0.0328±0.0009 | 0.0240±0.0007 | 3.14±0.11 | 3.15±0.10 | 0.9069±0.0054 |
| GRU-HLR | 14.61±0.52 | 7.84±0.39 | 24.83±1.32 | 14.61±0.61 | 0.9306±0.0076 | 0.0326±0.0012 | 0.0238±0.0008 | 3.11±0.11 | 3.12±0.11 | 0.9081±0.0072 |
| THLR | 15.99±0.53 | 8.51±0.29 | 25.95±0.56 | 15.97±0.67 | 0.9243±0.0033 | 0.0361±0.0017 | 0.0263±0.0011 | 3.46±0.17 | 3.46±0.17 | 0.8871±0.0107 |
| ROKAN | 34.64±2.25 | 19.95±1.98 | 57.85±4.79 | 60.33±40.32 | 0.6217±0.0646 | 0.1177±0.0378 | 0.0657±0.0118 | 10.04±3.04 | 8.72±1.52 | −0.3106±0.8825 |
| I-Mem | 14.69±0.21 | 7.87±0.15 | 25.12±0.59 | 14.67±0.44 | 0.9291±0.0033 | 0.0328±0.0006 | 0.0240±0.0004 | 3.14±0.06 | 3.16±0.06 | 0.9067±0.0032 |
| *Recent KT / memory* | | | | | | | | | | |
| EPsyINN | 14.22±0.25 | 7.65±0.19 | 24.63±0.83 | 14.02±0.19 | 0.9318±0.0046 | 0.0315±0.0005 | 0.0231±0.0004 | 3.00±0.05 | 3.01±0.05 | 0.9142±0.0026 |
| LefoKT | 16.22±0.52 | 8.62±0.31 | 26.41±1.27 | 16.44±0.45 | 0.9215±0.0076 | 0.0365±0.0012 | 0.0266±0.0009 | 3.50±0.12 | 3.52±0.11 | 0.8844±0.0080 |
| ReKT | 14.67±0.66 | 7.79±0.39 | 24.70±1.28 | 14.69±0.83 | 0.9313±0.0075 | 0.0327±0.0015 | 0.0239±0.0011 | 3.12±0.15 | 3.14±0.16 | 0.9073±0.0087 |
| LTKT | 14.48±0.21 | 7.70±0.18 | 24.68±0.84 | 14.38±0.24 | 0.9315±0.0047 | 0.0320±0.0005 | 0.0235±0.0004 | 3.06±0.05 | 3.07±0.05 | 0.9113±0.0027 |
| SPsyINN | 14.32±0.39 | 7.63±0.20 | 24.47±0.77 | 14.16±0.30 | 0.9327±0.0043 | 0.0318±0.0008 | 0.0234±0.0006 | 3.04±0.09 | 3.05±0.08 | 0.9121±0.0042 |
| KVFKT | 14.63±0.19 | 7.77±0.10 | 24.27±0.45 | 14.67±0.36 | 0.9338±0.0025 | 0.0325±0.0004 | 0.0238±0.0003 | 3.11±0.04 | 3.12±0.04 | 0.9086±0.0020 |
| GRKT | 15.28±0.69 | 8.11±0.27 | 25.03±0.93 | 15.26±0.62 | 0.9295±0.0053 | 0.0337±0.0012 | 0.0248±0.0011 | 3.24±0.14 | 3.24±0.13 | 0.9018±0.0070 |
| MIKT | 14.21±0.42 | 7.60±0.24 | 24.45±0.89 | 14.21±0.32 | 0.9327±0.0050 | 0.0317±0.0008 | 0.0231±0.0007 | 3.01±0.09 | 3.02±0.08 | 0.9131±0.0045 |
| MCKT | 14.46±0.25 | 7.77±0.23 | 24.63±1.08 | 14.55±0.40 | 0.9317±0.0061 | 0.0321±0.0007 | 0.0235±0.0005 | 3.07±0.07 | 3.09±0.07 | 0.9106±0.0037 |
| DGMKT | 14.42±0.37 | 7.67±0.20 | 23.97±0.48 | 14.53±0.49 | 0.9354±0.0026 | 0.0322±0.0009 | 0.0235±0.0006 | 3.07±0.09 | 3.09±0.10 | 0.9103±0.0050 |
| *Ours* | | | | | | | | | | |
| **OSR-HLR (Ours)** | **12.37±0.05** | **6.27±0.04** | **21.00±0.11** | **12.27±0.07** | **0.9504±0.0005** | **0.0274±0.0001** | **0.0199±0.0001** | **2.58±0.01** | **2.59±0.01** | **0.9348±0.0005** |
| **Improvement (%)** | **+12.96%** | **+17.55%** | **+12.38%** | **+12.51%** | **+1.61%** | **+12.83%** | **+13.73%** | **+14.01%** | **+13.91%** | **+2.26%** |

Table 19: shared_exact: Anki half-life and recall-probability metrics. Values are mean ± sample SD; lower is better except $R^2$.

| Method | Half-life | | | | | Recall probability | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | H-SMAPE ↓ | H-MAE ↓ | H-RMSE ↓ | H-MAPE ↓ | H-$R^2$ ↑ | P-RMSE ↓ | P-MAE ↓ | P-SMAPE ↓ | P-MAPE ↓ | P-$R^2$ ↑ |
| *Cognitive / formulaic* | | | | | | | | | | |
| Pimsleur | 39.42±0.02 | 26.09±0.43 | 432.16±9.92 | 764.10±13.06 | −3.2558±0.1957 | 0.1928±0.0004 | 0.1101±0.0002 | 29.33±0.06 | 19.18±0.02 | −0.1580±0.0045 |
| Leitner | 30.09±0.01 | 75.04±0.17 | 1369.06±2.13 | 1872.41±3.56 | −41.6901±0.1331 | 0.1731±0.0000 | 0.0987±0.0000 | 19.31±0.00 | 22.50±0.00 | 0.0669±0.0002 |
| ACT-R | 38.83±0.08 | 65.52±0.92 | 439.87±4.81 | 18339.93±276.41 | −3.4073±0.0963 | 0.1742±0.0001 | 0.1004±0.0001 | 19.32±0.02 | 21.76±0.02 | 0.0545±0.0015 |
| HLR | 38.17±0.14 | 18.67±0.36 | 186.24±5.46 | 3599.47±86.53 | 0.2094±0.0464 | 0.1722±0.0014 | 0.1012±0.0006 | 24.24±0.25 | 17.90±0.08 | 0.0759±0.0149 |
| DASH | 37.42±0.46 | 4.79±0.32 | 57.90±20.93 | 191.18±37.82 | 0.9147±0.0630 | 0.2419±0.0050 | 0.1326±0.0024 | 33.66±0.50 | 21.47±0.28 | −0.8237±0.0743 |
| DHP | 19.05±0.04 | 3.41±0.05 | 21.02±0.29 | 44.76±0.30 | 0.9899±0.0003 | 0.1269±0.0001 | 0.0748±0.0001 | 16.61±0.02 | 14.58±0.05 | 0.4985±0.0010 |
| FSRS | 19.90±0.35 | 3.32±0.15 | 24.37±1.58 | 49.65±4.76 | 0.9864±0.0018 | 0.1324±0.0007 | 0.0777±0.0004 | 17.27±0.09 | 16.46±0.08 | 0.4541±0.0054 |
| *Generic sequence* | | | | | | | | | | |
| RNN | 9.20±0.19 | 1.33±0.04 | 13.87±0.73 | 15.10±0.44 | 0.9956±0.0005 | 0.1175±0.0003 | 0.0620±0.0003 | 15.05±0.05 | 14.32±0.04 | 0.5700±0.0018 |
| GRU | 9.60±0.22 | 1.42±0.09 | 15.54±1.23 | 15.09±0.65 | 0.9945±0.0009 | 0.1180±0.0004 | 0.0627±0.0004 | 15.13±0.07 | 14.38±0.05 | 0.5666±0.0027 |
| LSTM | 9.65±0.31 | 1.45±0.08 | 15.65±0.86 | 15.13±0.73 | 0.9944±0.0006 | 0.1180±0.0003 | 0.0627±0.0004 | 15.13±0.06 | 14.41±0.08 | 0.5662±0.0024 |
| Transformer | 10.86±0.21 | 1.53±0.10 | 14.78±0.92 | 17.32±1.12 | 0.9950±0.0006 | 0.1201±0.0004 | 0.0648±0.0004 | 15.54±0.10 | 14.76±0.08 | 0.5504±0.0028 |
| DKT | 8.97±0.10 | 1.40±0.08 | 15.10±0.88 | 14.56±0.54 | 0.9948±0.0006 | 0.1173±0.0002 | 0.0616±0.0002 | 15.01±0.03 | 14.28±0.03 | 0.5715±0.0013 |
| *Memory / forgetting* | | | | | | | | | | |
| FIFAKT | 8.81±0.17 | 1.38±0.07 | 14.44±0.81 | 14.19±0.91 | 0.9952±0.0005 | 0.1170±0.0002 | 0.0612±0.0002 | 14.94±0.04 | 14.21±0.03 | 0.5735±0.0012 |
| GRU-HLR | 7.11±0.23 | 1.41±0.08 | 15.10±1.15 | 10.39±0.68 | 0.9948±0.0008 | 0.1046±0.0003 | 0.0543±0.0004 | 13.52±0.15 | 9.85±0.11 | 0.6593±0.0019 |
| THLR | 8.65±0.34 | 1.51±0.09 | 15.17±1.62 | 14.28±0.84 | 0.9947±0.0012 | 0.1071±0.0006 | 0.0570±0.0006 | 14.11±0.19 | 10.35±0.12 | 0.6425±0.0041 |
| ROKAN | 15.39±0.93 | 2.29±0.23 | 20.23±1.96 | 65.01±21.49 | 0.9906±0.0019 | 0.1306±0.0028 | 0.0722±0.0016 | 16.68±0.30 | 15.91±0.25 | 0.4686±0.0228 |
| I-Mem | 10.09±0.16 | 1.47±0.08 | 15.36±0.95 | 15.93±0.85 | 0.9946±0.0007 | 0.1185±0.0004 | 0.0635±0.0003 | 15.27±0.06 | 14.52±0.05 | 0.5626±0.0026 |
| *Recent KT / memory* | | | | | | | | | | |
| EPsyINN | 8.52±0.12 | 1.31±0.05 | 13.77±0.57 | 15.55±0.65 | 0.9957±0.0004 | 0.1166±0.0001 | 0.0605±0.0002 | 14.85±0.02 | 14.12±0.03 | 0.5768±0.0007 |
| LefoKT | 10.19±0.29 | 1.58±0.08 | 16.60±1.35 | 19.16±2.86 | 0.9937±0.0010 | 0.1188±0.0006 | 0.0635±0.0007 | 15.31±0.16 | 14.59±0.12 | 0.5602±0.0046 |
| ReKT | 9.47±0.11 | 1.60±0.10 | 16.73±1.39 | 20.37±1.64 | 0.9936±0.0011 | 0.1173±0.0002 | 0.0616±0.0002 | 14.98±0.03 | 14.34±0.04 | 0.5715±0.0014 |
| LTKT | 8.79±0.15 | 1.40±0.04 | 14.34±0.47 | 19.21±2.14 | 0.9953±0.0003 | 0.1168±0.0001 | 0.0607±0.0002 | 14.86±0.03 | 14.15±0.03 | 0.5752±0.0011 |
| SPsyINN | 8.40±0.08 | 1.33±0.06 | 14.42±0.83 | 15.75±0.87 | 0.9953±0.0005 | 0.1164±0.0001 | 0.0603±0.0001 | 14.81±0.01 | 14.09±0.02 | 0.5780±0.0004 |
| KVFKT | 9.36±0.11 | 1.52±0.05 | 15.30±0.49 | 20.46±0.74 | 0.9947±0.0003 | 0.1174±0.0001 | 0.0616±0.0002 | 14.98±0.03 | 14.28±0.04 | 0.5710±0.0010 |
| GRKT | 9.19±0.37 | 1.44±0.10 | 14.62±0.68 | 18.54±3.18 | 0.9951±0.0004 | 0.1170±0.0004 | 0.0612±0.0005 | 14.94±0.08 | 14.23±0.08 | 0.5732±0.0026 |
| MIKT | 8.74±0.20 | 1.44±0.08 | 15.44±0.77 | 17.74±1.38 | 0.9946±0.0005 | 0.1167±0.0002 | 0.0606±0.0003 | 14.85±0.03 | 14.13±0.04 | 0.5758±0.0012 |
| MCKT | 8.74±0.18 | 1.33±0.07 | 14.14±0.73 | 15.79±0.57 | 0.9954±0.0005 | 0.1167±0.0001 | 0.0608±0.0002 | 14.87±0.02 | 14.17±0.03 | 0.5760±0.0008 |
| DGMKT | 8.96±0.18 | 1.42±0.11 | 15.32±1.43 | 16.31±0.74 | 0.9946±0.0010 | 0.1168±0.0002 | 0.0610±0.0003 | 14.89±0.04 | 14.19±0.05 | 0.5748±0.0013 |
| *Ours* | | | | | | | | | | |
| **OSR-HLR (Ours)** | **4.79±0.04** | **0.72±0.01** | **7.88±0.20** | **5.22±0.08** | **0.9986±0.0001** | **0.1023±0.0000** | **0.0509±0.0001** | **12.80±0.01** | **9.20±0.02** | **0.6739±0.0003** |
| **Improvement (%)** | **+32.59%** | **+44.78%** | **+42.75%** | **+49.78%** | **+0.29%** | **+2.17%** | **+6.24%** | **+5.30%** | **+6.52%** | **+2.22%** |

Table 20: shared_exact: EdNet-Recall half-life and recall-probability metrics. Values are mean ± sample SD; lower is better except $R^2$.

| Method | Half-life | | | | | Recall probability | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | H-SMAPE ↓ | H-MAE ↓ | H-RMSE ↓ | H-MAPE ↓ | H-$R^2$ ↑ | P-RMSE ↓ | P-MAE ↓ | P-SMAPE ↓ | P-MAPE ↓ | P-$R^2$ ↑ |
| *Cognitive / formulaic* | | | | | | | | | | |
| Pimsleur | 158.68±14.73 | 287.66±68.42 | 467.51±143.97 | 99.39±3.67 | −0.5814±0.4390 | 0.4031±0.0771 | 0.3226±0.0900 | 84.81±30.00 | 58.01±11.04 | −3.8036±2.6724 |
| Leitner | 160.77±11.36 | 293.39±71.60 | 474.23±146.50 | 96.18±2.37 | −0.6161±0.4171 | 0.4049±0.0557 | 0.3232±0.0666 | 83.60±21.18 | 58.45±9.43 | −3.7181±1.9859 |
| ACT-R | 76.27±5.25 | 223.68±73.56 | 390.08±153.74 | 317.89±133.23 | 0.0439±0.1139 | 0.2925±0.0134 | 0.2263±0.0152 | 31.95±3.01 | 55.71±8.05 | −1.3634±0.4512 |
| HLR | 82.22±29.91 | 221.71±47.46 | 746.64±492.34 | 269.27±129.96 | −3.2473±3.7593 | 0.2987±0.0482 | 0.2289±0.0449 | 39.99±18.55 | 53.42±8.86 | −1.5822±1.2566 |
| DASH | 84.32±39.90 | 258.60±223.72 | 1003.04±2343.77 | 1674.04±5467.66 | −330.6153±1279.9400 | 0.3110±0.0802 | 0.2429±0.0828 | 45.16±31.46 | 54.40±10.90 | −1.9487±2.2642 |
| DHP | 81.50±19.27 | 226.34±48.08 | 373.54±104.30 | 350.47±154.57 | −0.3518±1.2117 | 0.2920±0.0122 | 0.2227±0.0143 | 35.13±5.12 | 53.79±7.66 | −1.3510±0.4128 |
| FSRS | 69.06±25.91 | 179.57±38.53 | 319.74±100.90 | 249.40±118.30 | 0.1512±0.5345 | 0.2813±0.0262 | 0.2166±0.0277 | 32.67±7.08 | 53.18±9.53 | −1.2201±0.6875 |
| *Generic sequence* | | | | | | | | | | |
| RNN | 55.47±10.44 | 157.72±49.66 | 281.01±117.26 | 168.29±55.01 | 0.4494±0.1887 | 0.2715±0.0147 | 0.2088±0.0165 | 30.04±3.01 | 47.79±6.82 | −1.0441±0.4503 |
| GRU | 59.09±13.64 | 161.69±52.70 | 283.70±116.39 | 174.58±61.46 | 0.4255±0.2177 | 0.2753±0.0181 | 0.2111±0.0179 | 31.03±3.51 | 50.39±6.88 | −1.0969±0.4513 |
| LSTM | 58.00±11.66 | 161.60±53.69 | 283.46±116.89 | 182.80±75.90 | 0.4399±0.1880 | 0.2737±0.0182 | 0.2102±0.0182 | 30.64±3.34 | 50.32±6.55 | −1.0745±0.4535 |
| Transformer | 61.29±10.52 | 188.35±65.76 | 326.88±126.87 | 269.71±125.47 | 0.2357±0.2808 | 0.2781±0.0163 | 0.2145±0.0176 | 30.52±3.00 | 53.20±8.49 | −1.1433±0.4717 |
| DKT | 55.36±11.38 | 154.80±49.48 | 275.06±110.38 | 180.06±70.24 | 0.4694±0.1750 | 0.2716±0.0156 | 0.2085±0.0171 | 29.92±3.09 | 52.01±8.55 | −1.0453±0.4561 |
| *Memory / forgetting* | | | | | | | | | | |
| FIFAKT | 53.85±11.30 | 158.02±53.49 | 286.22±117.71 | 177.25±114.96 | 0.4394±0.1662 | 0.2710±0.0174 | 0.2084±0.0195 | 29.87±3.04 | 51.82±8.94 | −1.0415±0.4934 |
| GRU-HLR | 54.63±10.02 | 151.48±44.24 | 270.96±107.46 | 166.84±71.61 | 0.4758±0.1900 | 0.2711±0.0167 | 0.2085±0.0186 | 29.97±3.15 | 50.00±7.41 | −1.0415±0.4770 |
| THLR | 56.01±9.64 | 158.57±52.55 | 278.54±103.71 | 219.18±109.00 | 0.4233±0.2499 | 0.2752±0.0172 | 0.2120±0.0198 | 30.16±2.99 | 52.70±8.95 | −1.1044±0.5028 |
| ROKAN | 64.77±11.07 | 202.67±70.12 | 350.59±121.94 | 281.68±115.09 | 0.0514±0.4649 | 0.2789±0.0156 | 0.2142±0.0163 | 31.12±3.38 | 53.04±8.88 | −1.1614±0.4994 |
| I-Mem | 60.42±13.78 | 167.94±46.07 | 311.25±108.30 | 194.81±93.27 | 0.2259±0.5144 | 0.2774±0.0277 | 0.2125±0.0255 | 31.46±6.53 | 52.25±8.78 | −1.1638±0.7237 |
| *Recent KT / memory* | | | | | | | | | | |
| EPsyINN | 55.34±10.19 | 153.09±45.67 | 269.50±104.85 | 192.11±62.25 | 0.4846±0.1855 | 0.2722±0.0158 | 0.2097±0.0183 | 29.81±3.06 | 52.06±8.26 | −1.0545±0.4588 |
| LefoKT | 61.57±13.87 | 180.50±40.92 | 299.50±68.69 | 305.45±167.08 | 0.0302±1.0288 | 0.2770±0.0198 | 0.2149±0.0236 | 30.36±3.44 | 53.30±9.19 | −1.1360±0.5472 |
| ReKT | 55.56±10.32 | 154.04±46.36 | 272.46±108.30 | 207.77±50.74 | 0.4760±0.1886 | 0.2724±0.0148 | 0.2099±0.0172 | 29.79±3.01 | 52.31±8.48 | −1.0563±0.4437 |
| LTKT | 56.32±11.18 | 157.82±40.97 | 279.10±92.77 | 232.92±111.54 | 0.3354±0.4643 | 0.2728±0.0166 | 0.2104±0.0191 | 29.86±3.11 | 52.41±8.82 | −1.0658±0.4790 |
| SPsyINN | 55.53±10.90 | 153.80±43.66 | 270.49±100.52 | 211.65±78.60 | 0.4506±0.2705 | 0.2727±0.0179 | 0.2105±0.0209 | 29.87±3.27 | 52.39±8.97 | −1.0690±0.5102 |
| KVFKT | 57.13±12.23 | 159.73±50.24 | 282.34±107.36 | 201.93±75.42 | 0.3909±0.3281 | 0.2735±0.0167 | 0.2107±0.0184 | 30.10±3.19 | 52.19±8.13 | −1.0723±0.4542 |
| GRKT | 58.64±14.11 | 162.14±55.14 | 283.94±121.67 | 224.05±66.00 | 0.4390±0.2035 | 0.2771±0.0200 | 0.2138±0.0215 | 30.77±3.78 | 52.88±8.82 | −1.1289±0.4936 |
| MIKT | 55.02±9.68 | 154.63±47.13 | 271.18±107.24 | 206.94±52.14 | 0.4866±0.1605 | 0.2719±0.0147 | 0.2094±0.0167 | 29.76±2.94 | 51.96±8.10 | −1.0464±0.4337 |
| MCKT | 55.27±10.18 | 155.43±44.28 | 277.82±103.52 | 219.12±89.83 | 0.4189±0.2867 | 0.2724±0.0165 | 0.2103±0.0195 | 29.79±3.08 | 52.32±8.85 | −1.0613±0.4824 |
| DGMKT | 55.42±9.58 | 153.24±47.28 | 274.96±105.43 | 207.90±64.90 | 0.4552±0.2109 | 0.2731±0.0175 | 0.2109±0.0213 | 29.93±3.12 | 52.47±8.87 | −1.0727±0.4994 |
| *Ours* | | | | | | | | | | |
| **OSR-HLR (Ours)** | **35.16±6.61** | **93.75±23.49** | **178.99±54.45** | **47.38±14.36** | **0.7536±0.1059** | **0.2573±0.0159** | **0.1953±0.0168** | **28.65±2.74** | **45.04±6.91** | **−0.8418±0.4498** |
| **Improvement (%)** | **+34.70%** | **+38.11%** | **+33.59%** | **+50.73%** | **+54.89%** | **+5.06%** | **+6.25%** | **+3.76%** | **+5.75%** | **+19.17%** |

Table 21: equal_postprocessor: MaiMemo half-life and recall-probability metrics. Values are mean ± sample SD; lower is better except $R^2$.

| Method | Half-life | | | | | Recall probability | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | H-SMAPE ↓ | H-MAE ↓ | H-RMSE ↓ | H-MAPE ↓ | H-$R^2$ ↑ | P-RMSE ↓ | P-MAE ↓ | P-SMAPE ↓ | P-MAPE ↓ | P-$R^2$ ↑ |
| *Cognitive / formulaic* | | | | | | | | | | |
| Pimsleur | 48.88±0.02 | 24.82±0.04 | 62.99±0.06 | 76.08±0.38 | 0.5543±0.0009 | 0.1365±0.0004 | 0.0858±0.0002 | 13.52±0.05 | 11.47±0.03 | −0.6131±0.0095 |
| Leitner | 42.73±0.00 | 17.19±0.01 | 42.37±0.02 | 50.39±0.07 | 0.7983±0.0002 | 0.1064±0.0000 | 0.0752±0.0000 | 9.72±0.00 | 10.69±0.00 | 0.0190±0.0006 |
| ACT-R | 45.81±0.02 | 27.15±0.16 | 95.69±1.06 | 345.39±2.99 | −0.0288±0.0227 | 0.0866±0.0001 | 0.0574±0.0000 | 7.23±0.00 | 7.85±0.00 | 0.3509±0.0010 |
| HLR | 44.67±0.03 | 20.40±0.09 | 51.55±0.29 | 159.14±1.15 | 0.7015±0.0033 | 0.0939±0.0005 | 0.0635±0.0002 | 9.28±0.06 | 8.67±0.03 | 0.2362±0.0083 |
| DASH | 32.86±0.03 | 15.52±0.10 | 45.02±0.48 | 36.46±0.18 | 0.7723±0.0049 | 0.0906±0.0002 | 0.0585±0.0001 | 8.82±0.03 | 8.06±0.01 | 0.2891±0.0039 |
| DHP | 22.89±0.01 | 11.87±0.01 | 35.85±0.14 | 24.07±0.17 | 0.8556±0.0011 | 0.0522±0.0000 | 0.0373±0.0000 | 4.85±0.01 | 4.91±0.00 | 0.7641±0.0002 |
| FSRS | 26.27±1.10 | 11.54±0.36 | 30.93±0.84 | 30.22±2.53 | 0.8924±0.0059 | 0.0616±0.0062 | 0.0431±0.0020 | 5.77±0.45 | 5.94±0.29 | 0.6687±0.0745 |
| *Generic sequence* | | | | | | | | | | |
| RNN | 14.03±0.24 | 7.56±0.20 | 24.34±0.94 | 13.92±0.27 | 0.9333±0.0052 | 0.0312±0.0005 | 0.0228±0.0004 | 2.97±0.05 | 2.97±0.05 | 0.9156±0.0027 |
| GRU | 13.73±0.13 | 7.30±0.08 | 23.58±0.40 | 13.61±0.09 | 0.9375±0.0021 | 0.0306±0.0002 | 0.0223±0.0002 | 2.90±0.03 | 2.91±0.03 | 0.9192±0.0012 |
| LSTM | 13.79±0.27 | 7.38±0.19 | 23.74±0.72 | 13.78±0.39 | 0.9366±0.0040 | 0.0307±0.0006 | 0.0224±0.0004 | 2.91±0.06 | 2.93±0.06 | 0.9181±0.0033 |
| Transformer | 14.72±0.35 | 8.02±0.35 | 25.39±1.51 | 14.79±0.37 | 0.9274±0.0090 | 0.0328±0.0007 | 0.0239±0.0005 | 3.13±0.07 | 3.14±0.07 | 0.9066±0.0040 |
| DKT | 13.73±0.12 | 7.34±0.11 | 23.62±0.59 | 13.73±0.13 | 0.9373±0.0032 | 0.0307±0.0003 | 0.0223±0.0002 | 2.91±0.03 | 2.93±0.04 | 0.9185±0.0016 |
| *Memory / forgetting* | | | | | | | | | | |
| FIFAKT | 14.03±0.23 | 7.61±0.21 | 24.47±0.83 | 14.03±0.24 | 0.9326±0.0046 | 0.0313±0.0005 | 0.0228±0.0004 | 2.97±0.06 | 2.98±0.05 | 0.9153±0.0028 |
| GRU-HLR | 13.94±0.36 | 7.44±0.28 | 23.83±0.96 | 13.90±0.41 | 0.9361±0.0053 | 0.0311±0.0009 | 0.0227±0.0006 | 2.95±0.08 | 2.96±0.08 | 0.9161±0.0049 |
| THLR | 14.80±0.35 | 7.90±0.19 | 24.63±0.42 | 14.77±0.43 | 0.9318±0.0023 | 0.0333±0.0011 | 0.0241±0.0007 | 3.16±0.11 | 3.16±0.10 | 0.9041±0.0063 |
| ROKAN | 30.79±2.24 | 18.41±1.89 | 55.37±5.76 | 53.37±36.03 | 0.6522±0.0750 | 0.1102±0.0395 | 0.0579±0.0118 | 8.90±3.05 | 7.65±1.52 | −0.1732±0.8787 |
| I-Mem | 13.94±0.14 | 7.44±0.10 | 23.95±0.39 | 13.89±0.23 | 0.9355±0.0021 | 0.0311±0.0003 | 0.0227±0.0002 | 2.96±0.03 | 2.97±0.03 | 0.9161±0.0016 |
| *Recent KT / memory* | | | | | | | | | | |
| EPsyINN | 13.66±0.14 | 7.30±0.12 | 23.65±0.54 | 13.51±0.10 | 0.9372±0.0029 | 0.0303±0.0003 | 0.0222±0.0002 | 2.88±0.02 | 2.89±0.02 | 0.9203±0.0014 |
| LefoKT | 14.89±0.35 | 7.92±0.22 | 24.78±0.85 | 14.99±0.34 | 0.9309±0.0047 | 0.0335±0.0010 | 0.0242±0.0006 | 3.17±0.08 | 3.18±0.07 | 0.9028±0.0057 |
| ReKT | 13.95±0.48 | 7.40±0.29 | 23.74±1.00 | 13.92±0.54 | 0.9366±0.0056 | 0.0311±0.0010 | 0.0227±0.0007 | 2.95±0.10 | 2.96±0.10 | 0.9160±0.0055 |
| LTKT | 13.81±0.13 | 7.34±0.13 | 23.78±0.60 | 13.72±0.15 | 0.9364±0.0032 | 0.0306±0.0003 | 0.0224±0.0002 | 2.91±0.03 | 2.92±0.03 | 0.9187±0.0016 |
| SPsyINN | 13.71±0.25 | 7.31±0.14 | 23.57±0.59 | 13.58±0.19 | 0.9375±0.0032 | 0.0305±0.0005 | 0.0223±0.0004 | 2.89±0.05 | 2.90±0.05 | 0.9194±0.0025 |
| KVFKT | 13.96±0.14 | 7.40±0.08 | 23.46±0.32 | 13.95±0.23 | 0.9382±0.0017 | 0.0310±0.0003 | 0.0227±0.0002 | 2.95±0.03 | 2.96±0.03 | 0.9167±0.0014 |
| GRKT | 14.39±0.45 | 7.64±0.20 | 24.00±0.68 | 14.36±0.42 | 0.9353±0.0037 | 0.0318±0.0008 | 0.0233±0.0006 | 3.03±0.08 | 3.04±0.08 | 0.9122±0.0043 |
| MIKT | 13.70±0.29 | 7.29±0.18 | 23.64±0.69 | 13.66±0.24 | 0.9371±0.0037 | 0.0306±0.0005 | 0.0223±0.0004 | 2.89±0.06 | 2.90±0.06 | 0.9190±0.0029 |
| MCKT | 13.78±0.15 | 7.35±0.14 | 23.61±0.72 | 13.76±0.23 | 0.9373±0.0039 | 0.0307±0.0004 | 0.0224±0.0003 | 2.91±0.04 | 2.92±0.04 | 0.9184±0.0020 |
| DGMKT | 13.78±0.21 | 7.31±0.12 | 23.24±0.37 | 13.78±0.24 | 0.9393±0.0019 | 0.0307±0.0004 | 0.0224±0.0003 | 2.91±0.04 | 2.93±0.05 | 0.9182±0.0024 |
| *Ours* | | | | | | | | | | |
| **OSR-HLR (Ours)** | **12.27±0.02** | **6.21±0.02** | **20.86±0.07** | **12.22±0.04** | **0.9511±0.0003** | **0.0273±0.0001** | **0.0198±0.0000** | **2.56±0.01** | **2.57±0.01** | **0.9353±0.0003** |
| **Improvement (%)** | **+10.18%** | **+14.79%** | **+10.26%** | **+9.58%** | **+1.26%** | **+9.94%** | **+10.79%** | **+11.00%** | **+10.87%** | **+1.64%** |

Table 22: equal_postprocessor: Anki half-life and recall-probability metrics. Values are mean ± sample SD; lower is better except $R^2$.

| Method | Half-life | | | | | Recall probability | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | H-SMAPE ↓ | H-MAE ↓ | H-RMSE ↓ | H-MAPE ↓ | H-$R^2$ ↑ | P-RMSE ↓ | P-MAE ↓ | P-SMAPE ↓ | P-MAPE ↓ | P-$R^2$ ↑ |
| *Cognitive / formulaic* | | | | | | | | | | |
| Pimsleur | 31.27±0.02 | 11.67±0.17 | 308.29±7.10 | 289.69±4.65 | −1.1658±0.0997 | 0.1708±0.0003 | 0.0981±0.0001 | 25.28±0.04 | 17.25±0.02 | 0.0918±0.0030 |
| Leitner | 22.00±0.01 | 25.37±0.04 | 627.52±0.54 | 464.06±0.55 | −7.9687±0.0154 | 0.1479±0.0000 | 0.0841±0.0000 | 17.72±0.00 | 18.87±0.00 | 0.3183±0.0002 |
| ACT-R | 33.03±0.07 | 14.61±0.13 | 156.88±0.92 | 4038.36±40.04 | 0.4394±0.0066 | 0.1582±0.0002 | 0.0912±0.0001 | 18.29±0.01 | 18.74±0.03 | 0.2204±0.0019 |
| HLR | 31.48±0.14 | 7.07±0.12 | 141.56±4.13 | 1139.48±29.60 | 0.5433±0.0267 | 0.1546±0.0011 | 0.0909±0.0005 | 22.03±0.21 | 16.26±0.08 | 0.2550±0.0111 |
| DASH | 30.11±0.57 | 3.14±0.25 | 51.77±22.03 | 131.46±38.00 | 0.9290±0.0602 | 0.2132±0.0063 | 0.1155±0.0028 | 29.34±0.60 | 19.07±0.32 | −0.4174±0.0820 |
| DHP | 12.87±0.03 | 2.09±0.02 | 15.19±0.13 | 24.89±0.14 | 0.9947±0.0001 | 0.1139±0.0000 | 0.0638±0.0000 | 14.93±0.01 | 11.93±0.03 | 0.5957±0.0003 |
| FSRS | 13.31±0.26 | 2.15±0.08 | 17.69±0.93 | 27.50±2.90 | 0.9929±0.0008 | 0.1143±0.0003 | 0.0646±0.0002 | 15.31±0.05 | 12.93±0.05 | 0.5932±0.0024 |
| *Generic sequence* | | | | | | | | | | |
| RNN | 7.08±0.10 | 1.15±0.02 | 12.42±0.31 | 9.35±0.19 | 0.9965±0.0002 | 0.1076±0.0001 | 0.0562±0.0001 | 14.16±0.03 | 11.77±0.02 | 0.6392±0.0007 |
| GRU | 7.23±0.12 | 1.20±0.05 | 13.51±0.79 | 9.43±0.27 | 0.9958±0.0005 | 0.1077±0.0001 | 0.0565±0.0002 | 14.18±0.03 | 11.77±0.03 | 0.6386±0.0010 |
| LSTM | 7.24±0.14 | 1.20±0.04 | 13.42±0.47 | 9.41±0.31 | 0.9959±0.0003 | 0.1078±0.0001 | 0.0564±0.0002 | 14.18±0.02 | 11.79±0.04 | 0.6382±0.0009 |
| Transformer | 7.86±0.12 | 1.23±0.02 | 12.67±0.21 | 10.51±0.34 | 0.9963±0.0001 | 0.1089±0.0002 | 0.0575±0.0002 | 14.40±0.05 | 11.99±0.05 | 0.6308±0.0014 |
| DKT | 6.91±0.07 | 1.17±0.04 | 12.95±0.40 | 9.04±0.16 | 0.9962±0.0002 | 0.1075±0.0001 | 0.0560±0.0001 | 14.14±0.01 | 11.75±0.01 | 0.6398±0.0005 |
| *Memory / forgetting* | | | | | | | | | | |
| FIFAKT | 6.88±0.08 | 1.17±0.04 | 12.75±0.46 | 8.95±0.30 | 0.9963±0.0003 | 0.1075±0.0001 | 0.0558±0.0001 | 14.12±0.02 | 11.72±0.02 | 0.6403±0.0005 |
| GRU-HLR | 5.88±0.10 | 1.17±0.03 | 12.79±0.48 | 7.40±0.28 | 0.9963±0.0003 | 0.1031±0.0001 | 0.0522±0.0001 | 13.08±0.04 | 9.38±0.04 | 0.6689±0.0007 |
| THLR | 6.69±0.16 | 1.24±0.04 | 13.08±0.62 | 9.26±0.32 | 0.9961±0.0004 | 0.1043±0.0002 | 0.0535±0.0003 | 13.33±0.05 | 9.63±0.05 | 0.6613±0.0016 |
| ROKAN | 11.32±0.70 | 1.74±0.12 | 16.98±1.35 | 26.06±6.65 | 0.9934±0.0011 | 0.1166±0.0021 | 0.0632±0.0012 | 15.26±0.21 | 12.88±0.20 | 0.5765±0.0154 |
| I-Mem | 7.43±0.10 | 1.22±0.04 | 13.26±0.40 | 9.76±0.33 | 0.9960±0.0002 | 0.1080±0.0002 | 0.0568±0.0001 | 14.25±0.04 | 11.84±0.04 | 0.6369±0.0010 |
| *Recent KT / memory* | | | | | | | | | | |
| EPsyINN | 6.73±0.04 | 1.13±0.02 | 12.17±0.21 | 10.03±0.29 | 0.9966±0.0001 | 0.1072±0.0001 | 0.0555±0.0000 | 14.08±0.01 | 11.67±0.01 | 0.6418±0.0005 |
| LefoKT | 7.48±0.12 | 1.25±0.03 | 13.59±0.50 | 12.00±2.02 | 0.9958±0.0003 | 0.1081±0.0003 | 0.0567±0.0003 | 14.27±0.07 | 11.88±0.06 | 0.6357±0.0017 |
| ReKT | 7.10±0.07 | 1.28±0.04 | 13.79±0.63 | 12.35±1.28 | 0.9957±0.0004 | 0.1072±0.0001 | 0.0558±0.0001 | 14.10±0.02 | 11.69±0.04 | 0.6418±0.0008 |
| LTKT | 6.87±0.09 | 1.17±0.03 | 12.46±0.21 | 12.36±1.62 | 0.9965±0.0001 | 0.1072±0.0001 | 0.0556±0.0001 | 14.07±0.01 | 11.65±0.02 | 0.6420±0.0006 |
| SPsyINN | 6.63±0.04 | 1.12±0.02 | 12.31±0.25 | 9.67±0.41 | 0.9965±0.0001 | 0.1071±0.0001 | 0.0554±0.0001 | 14.04±0.02 | 11.63±0.03 | 0.6430±0.0006 |
| KVFKT | 7.17±0.07 | 1.23±0.02 | 12.82±0.15 | 12.70±0.52 | 0.9963±0.0001 | 0.1075±0.0001 | 0.0560±0.0001 | 14.13±0.02 | 11.74±0.02 | 0.6401±0.0005 |
| GRKT | 7.10±0.22 | 1.19±0.06 | 12.57±0.32 | 11.70±2.03 | 0.9964±0.0002 | 0.1074±0.0002 | 0.0559±0.0003 | 14.11±0.05 | 11.71±0.05 | 0.6407±0.0015 |
| MIKT | 6.90±0.10 | 1.19±0.04 | 12.93±0.37 | 11.11±0.80 | 0.9962±0.0002 | 0.1072±0.0001 | 0.0556±0.0001 | 14.08±0.02 | 11.67±0.03 | 0.6417±0.0006 |
| MCKT | 6.75±0.07 | 1.12±0.02 | 12.27±0.14 | 9.61±0.34 | 0.9966±0.0001 | 0.1072±0.0001 | 0.0555±0.0001 | 14.07±0.02 | 11.65±0.03 | 0.6423±0.0006 |
| DGMKT | 6.85±0.11 | 1.19±0.04 | 13.13±0.57 | 9.90±0.49 | 0.9961±0.0003 | 0.1071±0.0001 | 0.0555±0.0002 | 14.06±0.02 | 11.64±0.03 | 0.6424±0.0007 |
| *Ours* | | | | | | | | | | |
| **OSR-HLR (Ours)** | **4.63±0.01** | **0.70±0.01** | **7.62±0.19** | **4.86±0.02** | **0.9987±0.0001** | **0.1020±0.0000** | **0.0507±0.0000** | **12.77±0.00** | **9.11±0.01** | **0.6759±0.0002** |
| **Improvement (%)** | **+21.27%** | **+37.33%** | **+37.43%** | **+34.39%** | **+0.21%** | **+1.06%** | **+2.90%** | **+2.42%** | **+2.88%** | **+1.05%** |

Table 23: equal_postprocessor: EdNet-Recall half-life and recall-probability metrics. Values are mean ± sample SD; lower is better except $R^2$.

| Method | Half-life | | | | | Recall probability | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | H-SMAPE ↓ | H-MAE ↓ | H-RMSE ↓ | H-MAPE ↓ | H-$R^2$ ↑ | P-RMSE ↓ | P-MAE ↓ | P-SMAPE ↓ | P-MAPE ↓ | P-$R^2$ ↑ |
| *Cognitive / formulaic* | | | | | | | | | | |
| Pimsleur | 153.93±16.83 | 281.31±65.55 | 458.39±139.88 | 96.58±2.58 | −0.5338±0.4611 | 0.3883±0.0813 | 0.3089±0.0933 | 80.12±30.96 | 56.61±11.49 | −3.5051±2.6979 |
| Leitner | 156.06±13.20 | 288.17±69.30 | 467.12±143.20 | 93.96±1.62 | −0.5781±0.4336 | 0.3895±0.0597 | 0.3085±0.0701 | 78.68±22.27 | 57.11±10.13 | −3.3988±2.0096 |
| ACT-R | 68.94±5.28 | 201.09±65.92 | 369.23±146.66 | 278.34±138.07 | 0.1429±0.1000 | 0.2870±0.0148 | 0.2215±0.0167 | 31.27±3.19 | 54.79±8.44 | −1.2810±0.4812 |
| HLR | 78.87±30.56 | 207.52±44.95 | 713.24±481.62 | 243.15±125.41 | −2.8720±3.4083 | 0.2949±0.0466 | 0.2257±0.0434 | 39.13±17.76 | 53.15±9.13 | −1.5141±1.2037 |
| DASH | 81.28±40.42 | 248.04±226.09 | 988.04±2347.65 | 1654.48±5472.21 | −330.5411±1279.9489 | 0.3070±0.0784 | 0.2392±0.0807 | 44.15±30.54 | 54.01±11.02 | −1.8691±2.1922 |
| DHP | 76.89±19.18 | 206.84±40.86 | 348.89±96.08 | 309.19±147.36 | −0.1979±1.1091 | 0.2883±0.0115 | 0.2195±0.0137 | 34.29±4.69 | 53.29±7.79 | −1.2928±0.3996 |
| FSRS | 66.09±25.30 | 168.59±38.23 | 303.46±99.27 | 226.73±116.97 | 0.2295±0.5058 | 0.2789±0.0249 | 0.2146±0.0266 | 32.17±6.53 | 52.85±9.48 | −1.1810±0.6595 |
| *Generic sequence* | | | | | | | | | | |
| RNN | 52.72±9.38 | 147.45±47.38 | 266.96±112.96 | 153.18±55.54 | 0.5097±0.1580 | 0.2704±0.0148 | 0.2077±0.0167 | 29.86±2.90 | 47.60±6.87 | −1.0276±0.4487 |
| GRU | 56.36±12.74 | 152.40±50.82 | 272.42±114.18 | 159.23±61.55 | 0.4796±0.1844 | 0.2736±0.0176 | 0.2096±0.0177 | 30.76±3.27 | 50.11±6.89 | −1.0721±0.4472 |
| LSTM | 55.19±11.13 | 151.44±51.92 | 270.65±114.45 | 166.88±75.84 | 0.4965±0.1625 | 0.2724±0.0177 | 0.2090±0.0180 | 30.41±3.15 | 50.09±6.54 | −1.0550±0.4500 |
| Transformer | 58.16±10.29 | 172.45±59.34 | 305.39±122.00 | 240.95±123.55 | 0.3237±0.2725 | 0.2756±0.0166 | 0.2121±0.0178 | 30.30±3.09 | 52.73±8.61 | −1.1087±0.4845 |
| DKT | 52.92±10.69 | 145.59±48.22 | 262.39±107.98 | 167.28±70.60 | 0.5229±0.1513 | 0.2705±0.0156 | 0.2075±0.0173 | 29.77±2.99 | 51.83±8.62 | −1.0301±0.4570 |
| *Memory / forgetting* | | | | | | | | | | |
| FIFAKT | 51.47±11.09 | 147.92±51.43 | 271.68±113.85 | 163.85±106.73 | 0.4955±0.1529 | 0.2698±0.0174 | 0.2072±0.0195 | 29.72±3.04 | 51.60±8.94 | −1.0242±0.4934 |
| GRU-HLR | 52.24±9.08 | 142.74±43.36 | 259.62±104.96 | 153.80±71.90 | 0.5270±0.1572 | 0.2702±0.0165 | 0.2076±0.0187 | 29.82±3.04 | 49.83±7.40 | −1.0269±0.4737 |
| THLR | 53.88±9.85 | 147.67±47.38 | 262.51±100.23 | 197.71±103.63 | 0.4882±0.2249 | 0.2736±0.0172 | 0.2103±0.0195 | 30.06±3.02 | 52.35±8.93 | −1.0816±0.5023 |
| ROKAN | 60.92±11.07 | 184.93±66.90 | 325.49±120.17 | 252.93±113.98 | 0.1850±0.4101 | 0.2765±0.0164 | 0.2121±0.0170 | 30.78±3.40 | 52.66±8.97 | −1.1268±0.5100 |
| I-Mem | 57.38±13.31 | 156.53±45.74 | 294.67±105.22 | 178.89±93.42 | 0.3044±0.4735 | 0.2760±0.0274 | 0.2113±0.0254 | 31.23±6.34 | 52.03±8.80 | −1.1427±0.7146 |
| *Recent KT / memory* | | | | | | | | | | |
| EPsyINN | 52.73±9.53 | 143.09±44.31 | 256.83±102.55 | 174.64±62.27 | 0.5359±0.1603 | 0.2709±0.0157 | 0.2085±0.0182 | 29.66±2.98 | 51.80±8.27 | −1.0354±0.4569 |
| LefoKT | 58.00±13.14 | 162.37±39.79 | 275.78±73.43 | 269.45±153.97 | 0.2162±0.7791 | 0.2751±0.0197 | 0.2130±0.0234 | 30.14±3.36 | 52.91±9.24 | −1.1079±0.5453 |
| ReKT | 52.91±9.54 | 144.00±44.81 | 259.51±105.08 | 189.06±55.30 | 0.5279±0.1631 | 0.2711±0.0148 | 0.2086±0.0173 | 29.65±2.94 | 52.05±8.50 | −1.0366±0.4433 |
| LTKT | 53.42±10.50 | 147.00±40.23 | 264.48±93.01 | 210.72±106.21 | 0.4287±0.3541 | 0.2715±0.0164 | 0.2090±0.0189 | 29.67±3.00 | 52.15±8.77 | −1.0459±0.4734 |
| SPsyINN | 53.05±10.48 | 143.79±42.56 | 257.56±98.82 | 192.29±77.58 | 0.5089±0.2243 | 0.2716±0.0177 | 0.2094±0.0208 | 29.73±3.20 | 52.17±8.94 | −1.0519±0.5051 |
| KVFKT | 54.45±11.39 | 148.83±48.20 | 268.09±105.35 | 181.58±72.01 | 0.4636±0.2605 | 0.2720±0.0164 | 0.2092±0.0184 | 29.92±3.07 | 51.90±8.17 | −1.0505±0.4524 |
| GRKT | 55.77±13.28 | 151.33±53.15 | 270.91±118.33 | 202.38±66.55 | 0.4930±0.1819 | 0.2754±0.0195 | 0.2122±0.0212 | 30.54±3.60 | 52.59±8.81 | −1.1039±0.4885 |
| MIKT | 52.36±9.01 | 144.13±45.37 | 257.99±104.43 | 187.35±54.17 | 0.5369±0.1465 | 0.2706±0.0147 | 0.2082±0.0168 | 29.62±2.88 | 51.71±8.11 | −1.0284±0.4339 |
| MCKT | 52.86±9.84 | 145.40±43.07 | 263.42±101.50 | 198.70±87.23 | 0.4857±0.2368 | 0.2712±0.0164 | 0.2090±0.0193 | 29.65±3.02 | 52.08±8.80 | −1.0425±0.4778 |
| DGMKT | 53.34±9.61 | 144.33±44.89 | 262.66±102.06 | 190.19±67.93 | 0.5008±0.1993 | 0.2719±0.0173 | 0.2097±0.0209 | 29.79±3.08 | 52.23±8.83 | −1.0550±0.4957 |
| *Ours* | | | | | | | | | | |
| **OSR-HLR (Ours)** | **34.25±6.71** | **91.00±23.22** | **175.10±55.06** | **44.88±13.99** | **0.7629±0.1072** | **0.2569±0.0159** | **0.1950±0.0167** | **28.61±2.71** | **44.94±6.90** | **−0.8371±0.4485** |
| **Improvement (%)** | **+33.46%** | **+36.25%** | **+31.82%** | **+52.23%** | **+42.09%** | **+4.77%** | **+5.90%** | **+3.41%** | **+5.58%** | **+18.27%** |

Table 24: Fixed-predictor comparison against six reference schedulers and three component ablations. Values are Test means ± sample SD under fixed OSR-HLR.

| Scheduler | Retention AUC ↑ | Time below target ↓ | Final recall ↑ | Reviews/deck |
|---|---|---|---|---|
| *Complete policy* | | | | |
| **OSR-ACRI (Ours)** | 0.5632 ± 0.2042 | 0.6501 ± 0.1953 | 0.5667 ± 0.2612 | 11680.0 ± 0.0 |
| *Reference schedulers* | | | | |
| Threshold | 0.5071 ± 0.1563 | 0.7027 ± 0.1601 | 0.4730 ± 0.1816 | 11680.0 ± 0.0 |
| Half-Life | 0.3398 ± 0.0795 | 0.9252 ± 0.0137 | 0.2637 ± 0.0761 | 7764.3 ± 3237.6 |
| MEMORIZE | 0.4860 ± 0.1607 | 0.6933 ± 0.1500 | 0.4359 ± 0.1902 | 11680.0 ± 0.0 |
| SSP-MMC | 0.4981 ± 0.1503 | 0.6958 ± 0.1463 | 0.4525 ± 0.1733 | 11680.0 ± 0.0 |
| ANKI | 0.4963 ± 0.1544 | 0.7894 ± 0.1157 | 0.4635 ± 0.1830 | 11680.0 ± 0.0 |
| RANDOM | 0.4318 ± 0.1509 | 0.8550 ± 0.0755 | 0.3504 ± 0.1452 | 10187.8 ± 2667.1 |
| *Component ablations* | | | | |
| No delay index | 0.5071 ± 0.1563 | 0.7011 ± 0.1591 | 0.4731 ± 0.1794 | 11680.0 ± 0.0 |
| No calibration | 0.5594 ± 0.2076 | 0.6623 ± 0.2020 | 0.5686 ± 0.2621 | 11680.0 ± 0.0 |
| No queue age | 0.3890 ± 0.1083 | 0.7514 ± 0.1362 | 0.3825 ± 0.1808 | 11680.0 ± 0.0 |

**Dataset domains.** The three benchmarks cover vocabulary learning (MaiMemo), general-purpose flashcards (Anki), and exam-oriented question practice (EdNet-Recall). They span two native spaced-repetition logs and one adaptive-practice log, but other domains, such as procedural skills, mathematics, or clinical training, may show different forgetting dynamics and remain to be tested.

**Log granularity.** The public releases provide review outcomes, intervals, and ratings, but not response latency, item content, or study context. OSR-HLR uses none of these signals, so its gains do not depend on them. Logs that include such signals could further improve half-life estimates.

**Label support.** Half-life labels are fitted from aggregated recall rates, so each memory state needs enough reviews at several query delays. Following prior half-life construction, we keep states with sufficient support. Rarely reviewed histories are therefore underrepresented in the evaluation, as in earlier work on these datasets.

**Simulation horizon.** The scheduling studies run for 365 days. This covers the typical span of a study course, but longer-term retention over several years is not evaluated.

**Future work.** The next step is prospective evaluation: a randomized learner A/B test would measure realized retention, workload, and heterogeneous effects under operational capacity constraints, and would test the transition assumptions shared by offline simulators. Further directions include extending the benchmark to additional learning domains, incorporating item content when it is available, and evaluating transfer across platforms.

## K Reproducibility Statement

We document every component needed to reproduce the reported results. Section 3 defines the prediction task and the half-life labels, and Section 4 specifies OSR-HLR and OSR-ACRI. Appendix E lists all input features, and Appendix F gives the data construction, split procedure, label-visibility boundary, optimization settings, and baseline adaptations. Appendix G describes the scheduling simulators, and Appendix H defines all metrics and statistical units.

All three datasets are public. MaiMemo is released under CC BY-NC 4.0, the Anki review logs are available through a gated public release, and EdNet-KT1 is a public benchmark. We do not redistribute raw logs; instead, the processed datasets are reconstructed from the original sources with deterministic, outcome-independent split assignments. Every reported number is computed from machine-readable result files with fixed random seeds (42–51 for MaiMemo and Anki, 42–44 for EdNet-Recall), and all hyperparameters are selected on Validation data only.

Upon acceptance, we will publicly release the complete source code under an open-source license. The release will cover data preprocessing, model training and evaluation for OSR-HLR and all baselines, the OSR-ACRI scheduler and simulators, and the scripts that regenerate every table and figure, together with split identifiers, frozen protocol files, configuration files, and trained checkpoints.