MINI-PSYCH-AGENT: EVALUATING HUMAN-LIKE BEHAVIOR IN LLM AGENTS THROUGH INTERACTIVE PSYCHOLOGICAL ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

A Large Language Model (LLM), its harness, and the test-time compute budget jointly determine its behavior. Consequently, the same underlying model, embedded in different agentic workflows or allocated different amounts of test-time compute, can exhibit markedly different performance. Psychologists face an analogous challenge when evaluating human performance: the same person may behave differently under different incentives and levels of deliberation. Psychological tasks therefore provide an ideal testbed for studying such behavioral variation. However, when adapted to LLMs, these tasks are typically text-transcribed, replaying fixed, language-rendered experiments as static prompts. This format is ill-suited to agent evaluation, where actions can alter future observations and task states. Conversely, although dynamic environments are prevalent in the evaluation of LLM agents, they rarely ask whether the resulting behavior is human-like. Here, we bridge this gap by transforming psychological tasks into dynamic environments in which agents can interact and by using a learned human-policy model to evaluate the human-likeness of the resulting behavioral trajectories. Across these environments, we find that the human-likeness of a given LLM varies with its agentic configuration and the psychological task. These findings motivate a more holistic evaluation of LLM agents in psychology-inspired environments, accounting for both the backbone model and its agentic scaffold.

1 INTRODUCTION

Psychological tasks are originally designed to isolate and test specific aspects of human cognition. To understand aspects of machine cognition, it is therefore natural for psychologists to adapt their carefully designed tasks to examine the behavior of LLMs. However, a common approach to evaluating LLMs on psychological tasks is to replay, in natural language, the experimental experiences originally encountered by human participants (Binz & Schulz, 2023; Binz et al., 2025; 2026; Bai et al., 2025; Frank & Goodman, 2026; Liu et al., 2025; Zhu & Griffiths, 2024). For example, the Centaur model, which is a finetuned Llama-3.1-70B model using psychological data, transcribes each experiment into a text sequence containing the task instructions, stimuli, and trial-by-trial interaction history (Binz et al., 2025). At each response point, the LLM is conditioned on the participant's experience up to, but not including, the response at that trial and evaluated according to the likelihood it assigns to the participant's observed response.

Verbatim replay provides a controlled and statistically principled measure of human-likeness through next-response prediction, following a long tradition of cognitive modeling (Daw, 2011; Wilson & Collins, 2019). But it does not directly assess closed-loop behavior, in which an agent's actions determine its subsequent observations and task states. As a result, replaying a human participant's experience evaluates model behavior only under the distribution of task states and histories generated by the participant (Binz et al., 2025; Xie & Zhu, 2025). This can conceal compounding errors: at every trial, the model is always returned to the participant's recorded history and therefore never has to recover from the consequences of its own earlier non-human-like decisions.

But an autonomous agent may take an unusual action and enter a state that few or no human participants encountered. Replay provides no evidence about its behavior in such states. The supplied

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.

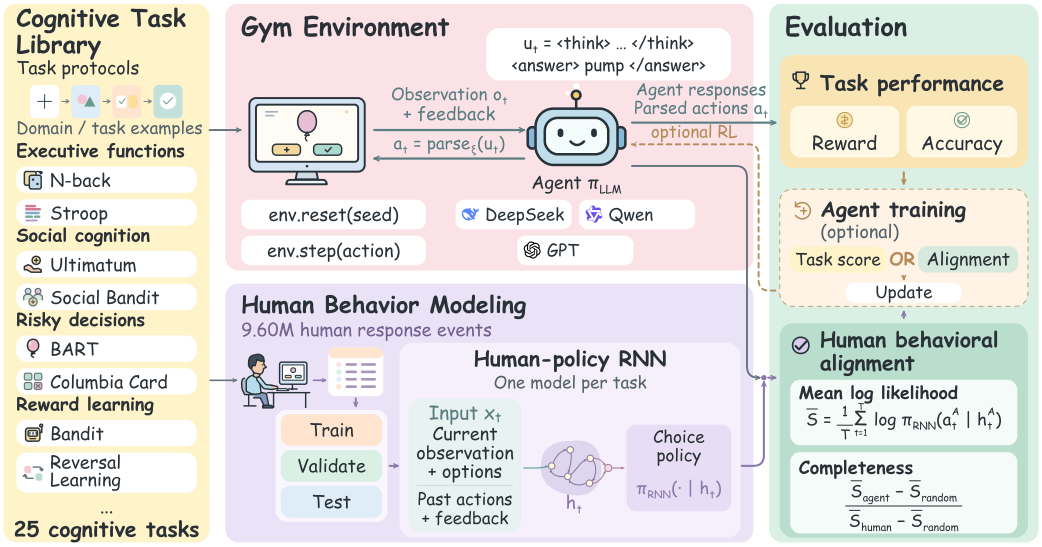

Figure 1: Overview of `MindGym`. LLM agents interact with 25 psychological tasks through a Gym-style interface. Their trajectories are evaluated for task performance and alignment with task-specific RNN models trained on human behavior. Displayed tasks and backbones are illustrative. No training or fine-tuning of LLMs were performed in this study; the dashed training loop illustrates a possible extension only.

human history may also be unlikely under the model's own policy. For example, the replay transcript may state that "you previously chose option A," even though the model would almost certainly have chosen option B. The model is thus asked to predict a response conditional on a history that it would rarely generate through its own decisions.

Evaluating autonomous agents therefore also requires assessing behavior under the histories generated by their own actions. To support this evaluation, we developed `MindGym`, a suite of interactive environments comprising 25 distinct psychological tasks that probe executive functions (e.g., the N-back task), social cognition (e.g., the ultimatum game), risky choice (e.g., the Balloon Analogue Risk Task), and reward learning (e.g., bandit tasks). Unlike approaches that only transcribe participants' experiences into text, we transform these tasks into interactive environments in which both humans and LLM agents can act through a defined set of permissible actions. As a result, these environments allow entirely new action sequences to be generated and evaluated under instructions similar to those given to the original human participants.

`MindGym` thus offers a unique opportunity to examine the behavior of LLM agents and evaluate their workflow designs through psychology-inspired environments. We created five `mini-psych-agent` variants by varying how agents maintain and use an external memory file (i.e., `MEMORY.md`). These variants range from a stateless baseline without persistent memory to more sophisticated approaches to context management, including retrieval-augmented generation (RAG) and surprisal-based episodic memory. External memory provides a practical mechanism for maintaining persistent agent state across model calls, allowing information from earlier interactions to be stored, updated, and retrieved to guide subsequent behavior.

Evaluating the human-likeness of `mini-psych-agent` variants in `MindGym` introduces an additional challenge: how to assess human-likeness when few or no matching human behavioral sequences are available. Our agents can generate novel trajectories that no human participant in the dataset has experienced. We address this challenge by training recurrent neural networks (RNNs) on human behavioral sequences, with one RNN for each of the 25 psychological tasks. Given sufficient human data, these RNNs can generalize to unseen histories within each task and serve as learned evaluators of agent behavior. We then freeze the RNNs and assess human-likeness using the likelihood these learned human-policy models assign to agent-generated actions, conditioned on the corresponding observations and interaction histories.

Our contributions are threefold. (i) We standardized and abstracted a range of psychological experiments into interactive environments in which agents can act and receive feedback. (ii) We developed novel likelihood metrics for evaluating the human-likeness of behavioral sequences using RNNs trained to model human policies. (iii) We systematically evaluated different memory-management strategies for LLM agents across interactive psychological environments.

2 Related Work

Evaluating human alignment of LLMs. Psychological experiments are now widely adapted to characterize LLM behavior, often through static prompts such as vignettes, questionnaires, and choice problems (e.g., Binz & Schulz, 2023; Zhu & Griffiths, 2024; Bai et al., 2025; Liu et al., 2025). For example, CogGym standardizes 258 commonsense reasoning experiments and compares 50 LLMs with humans on matched trials (Ying et al., 2026). However, fewer studies evaluate LLMs in closed-loop interactions, where the model's own choices determine its subsequent task states (e.g., Binz & Schulz, 2023; Coda-Forno et al., 2024; Schubert et al., 2024; Akata et al., 2025). These studies often summarize behavior using a few task-specific signatures, such as regret, stay probabilities, cooperation rates, or fitted model parameters (e.g., CogBench (Coda-Forno et al., 2024) and CogArena (Fernandez et al., 2025)). Although informative, these summaries require bespoke analyses for each task and retain only selected aspects of trial-level behavior. In parallel, interactive environments are well established for training and evaluating non-LLM agents. Psychlab implements psychological experiments for both humans and deep RL agents (Leibo et al., 2018), while NeuroGym provides neuroscience tasks for training neural networks (Molano-Mazón et al., 2022). For LLM agents, game-based benchmarks such as BALROG and TextArena provide interactive environments for evaluating task performance (Paglieri et al., 2025; Guertler et al., 2025). However, these benchmarks primarily assess game progress or skill ratings rather than the human-likeness of the LLM behavior. A remaining challenge is therefore to combine interactive environments with a common measure of human-likeness for trajectories generated by LLM agents.

Learned models of human behavior. Psychologists commonly develop computational models to explain cognitive mechanisms or predict human behavior. Whether interpretable cognitive models or black-box neural networks, these models are typically themselves the object of evaluation. For sequential tasks, a common criterion is how well they predict each response from a participant's preceding history (Daw, 2011; Wilson & Collins, 2019). Flexible learned models increasingly outperform hand-crafted theories in risky and strategic choice (Peterson et al., 2021; Zhu et al., 2025) and sequential decision-making (Dezfouli et al., 2019; Ji-An et al., 2025), while hybrid neural-cognitive models retain interpretable structure (Eckstein et al., 2026; Cao & Eckstein, 2025). In sufficiently data-rich settings, neural networks can also approach empirically estimated predictive limits (Kuperwajs et al., 2023; Yan et al., 2026), capturing maximal explainable variances in human data. Learned neural network models have also guided cognitive theory refinement through scientific regret minimization (Agrawal et al., 2020) and provided behavioral representations for evaluating LLM-generated behavior (Lei et al., 2026). In this paper, we use RNNs trained on human trajectories in each psychological task to evaluate the human-likeness of trajectories generated by autonomous LLM agents in the corresponding interactive environment. This extends the role of learned models from objects of evaluation to evaluators of novel agent-generated behavior within the same task.

Memory management in LLM agents. An LLM agent's behavior depends not only on its backbone but also on its scaffold (Yang et al., 2024; Kapoor et al., 2024), of which memory management is a central component. CoALA frames LLM agents as cognitive architectures with working and long-term memory (Sumers et al., 2024), while concrete designs include RAG (Lewis et al., 2020), memory streams with retrieval and reflection (Park et al., 2023), OS-inspired context management (Packer et al., 2023), and surprise-based event segmentation (Fountas et al., 2025). Memory benchmarks, however, commonly emphasize task performance, including question answering over long interaction histories and, more recently, on-policy conversational evaluation (Wu et al., 2025; Cheng et al., 2026). Memory management also shapes decisions: in a study of multi-armed bandits, satisfactory exploration emerged only when GPT-4 combined chain-of-thought reasoning with an externally summarized interaction history (Krishnamurthy et al., 2024). Prior work has also examined how memory contributes to behavioral believability (Park et al., 2023) and persona-consistent decisions (Peng et al., 2026). To our knowledge, however, no prior work has systematically examined how different memory-management strategies affect the human-likeness of agent-generated trajec-

tories across a suite of interactive psychological tasks. We address this gap by testing whether a memory design elicits more human-like behavior and how this relationship varies across tasks.

3 Evaluation Framework

Interactive psychological environments. Following the agent-environment interface used by OpenAI Gym (Brockman et al., 2016), we implement the `MindGym` environments using Gymnasium (Towers et al., 2024). Each environment supports interaction through natural-language instructions and observations, together with structured, task-specific actions. Specifically, each environment maintains an internal state that evolves in response to the agent's actions. As illustrated in Figure 1, interaction under task configuration ξ proceeds as

$$u_t \sim \pi_{\text{LLM}}(\cdot \mid c_t), \;\; a_t = \text{parse}_\xi(u_t), \;\; s_{t+1} \sim P_\xi(\cdot \mid s_t, a_t), \;\; o_{t+1} = g_\xi(s_{t+1}). \tag{1}$$

where t denotes the decision step and c_t is the input context constructed by the agent runtime. This context combines the task instructions and current textual observation o_t with any memory or interaction history supplied by the agent architecture (see Equation 2 for details). Conditioned on c_t, the LLM generates its response u_t under policy π, which is further parsed into an executable action a_t, which updates the internal environment state s_t through the transition kernel P_ξ. The rendering function g_ξ exposes participant-visible information and feedback permitted by the task protocol, while internal scoring and diagnostic information are recorded separately. Repeated interaction produces a behavioral trajectory shaped by the agent's own choices and their consequences.

The present study evaluates 25 psychological tasks spanning executive control, feedback learning and adaptation, information acquisition, social learning and strategic interaction, risk and decision making, planning and transfer, memory retrieval and integration, and affective updating. These categories reflect the primary demands of the implemented tasks. We draw on human behavioral datasets from Psych-201 (Binz et al., 2026) and datasets identified through OpenCogData (Zorowitz & NIMH DSST, 2024). We selected tasks with human behavioral data from more than 500 participants to support reliable training and held-out evaluation of task-specific RNN models of human behavior. For each task, we draw on published paradigms and available experimental materials to specify the available actions, outcome dynamics, feedback schedule, and stopping conditions. The full task inventory and adaptations required for language-based interaction are documented in Appendix A.

LLM agents and memory configurations. We implement `mini-psych-agent` as a lightweight, model-agnostic harness that combines an LLM with explicit memory management and an execution loop. At decision step t under task ξ, the runtime constructs the input context, obtains an LLM response, validates the proposed decision, and updates memory after execution:

$$c_t = C(I_\xi, o_t, m_t^\phi, \mathcal{A}_t), \;\; u_t \sim \pi_{\text{LLM}}(\cdot \mid c_t), \;\; a_t = \text{parse}_\xi(u_t), \;\; M_{t+1} = U_\phi(M_t, e_t). \tag{2}$$

where $C(\cdot)$ constructs the context, and U_ϕ updates memory according to memory-management strategy ϕ. Here, I_ξ denotes the task instructions, M_t the stored memory state, m_t^ϕ the memory content supplied to the LLM under strategy ϕ, and $\mathcal{A}_t$ defines the set of valid decisions. The tuple $e_t = (o_t, a_t, o_{t+1})$ records the completed, participant-visible transition. Only validated actions are executed. Memory is reset between runs (i.e., $M_0 = \varnothing$), and the LLM's parameters remain fixed.

Let $H_t = (e_0, \ldots, e_{t-1})$ denote the recorded public interaction history. We compare five configurations that differ in the memory content supplied to the LLM:

$$m_t^\phi = \begin{cases} \varnothing, & \text{Stateless}, \\ H_t, & \text{Cumulative}, \\ \text{Last}_6(H_t), & \text{Window}, \\ \text{Top}_6(H_t; \rho_t), & \text{Retrieval}, \\ (\sigma_1, \ldots, \sigma_{J_t}) \,\|\, B_t, & \text{Surprise}. \end{cases} \tag{3}$$

where $\|$ denotes concatenation. All configurations also receive the current observation o_t and task instructions I_ξ. The Stateless agent therefore receives no additional memory beyond these inputs. On the other extreme, the Cumulative agent always receives its full recorded interaction history at every decision step.

Between the Stateless and Cumulative extremes, we consider more selective memory strategies inspired by cognitive psychology (Miller, 1956; Tulving & Thomson, 1973). The Window agent receives up to six most recent interaction records. The Retrieval agent instead receives up to six records with the highest relevance scores, $\rho_t(e_i) = \cos(\mathrm{tf}(o_t), \mathrm{tf}(o_i))$, where tf denotes word-count vectors and ties favor more recent records.

Finally, inspired by surprise-based event segmentation in EM-LLM (Fountas et al., 2025), our Surprise agent uses elicited forecasts of public outcomes or stimuli to partition experience into events. Each closed event E_j is replaced by a textual summary $\sigma_j = \mathrm{Summarize}_{\mathrm{LLM}}(E_j)$. At decision step t, the agent receives the J_t completed summaries in chronological order, followed by the current event's uncompressed records B_t.

Learned human-policy model. To assess human-like behavior in states visited by few or no human participants, we train a separate RNN on human sequences for each of the 25 psychological tasks. Each model is trained to approximate the human policy conditioned on the current task context and preceding interaction history, providing a task-specific behavioral reference:

$$\pi_{\mathrm{RNN}}(a_t \mid x_{1:t}) \approx \pi_{\mathrm{human}}(a_t \mid x_{1:t}), \tag{4}$$

where $x_{1:t}$ denotes the sequence of structured inputs available before action a_t is selected, including the current stimuli or options and preceding responses and feedback. That is, $x_{1:t}$ is an abstraction of c_t in Equation 2.

Crucially, our RNNs use symbolic encodings of task features and responses, following common practice in neural network models of human behavior (Ji-An et al., 2025; Eckstein et al., 2026; Cao & Eckstein, 2025; Kuperwajs et al., 2023; Yan et al., 2026). Unlike LLM-based approaches that represent psychological tasks in natural language (Binz & Schulz, 2023), our RNN models learn directly from these structured representations. They therefore operate on task abstractions rather than directly processing the stimuli presented to participants, as "stimulus-computable" models do (Frank & Goodman, 2026). This abstraction focuses learning on task-relevant behavioral dependencies, with the aim of supporting generalization to novel interaction histories within the same task.

Training task-specific RNNs. For each task, we use a single-layer gated recurrent unit (GRU) (Cho et al., 2014) to maintain a recurrent state and generate task-specific predictions:

$$h_t = \mathrm{GRU}\big(h_{t-1}, x_t\big), \;\; a_t = g(h_t), \tag{5}$$

where x_t encodes the task inputs available before response a_t, h_t is the hidden state of GRU, and $g(\cdot)$ produces predictions a_t, comprising categorical action probabilities, continuous predictions, or both, depending on the psychological task. Recurrent states are reset at task-defined sequence boundaries.

We trained these RNN models through next-action prediction with teacher forcing: each human response is predicted from the current task inputs and preceding observed human responses and feedback. Categorical decisions are trained with cross-entropy, and continuous decisions with task-specific regression losses; tasks with multiple response types combine the corresponding objectives. Human data are partitioned approximately 80/10/10 into training, validation, and test sets by source-defined identity groups, with each group assigned to a single partition. Checkpoints are selected based on validation performance and frozen for subsequent evaluation. Task-specific encodings, losses, and training configurations are detailed in Appendix B.

Predictive validation. On held-out human trajectories, the RNNs exceeded uniform-random prediction accuracy by 36.3 ± 13.4 percentage points (mean $\pm$ SD across tasks), with above-chance performance on all 23 tasks with categorical decisions (Figure 2a). The RNNs also capture qualitative patterns in human behavior, including behavioral adaptation following reversals in the reversal learning task (see Figure 2b), declining accuracy with increasing N-back load (see Figure 2c), and increasing preference for advantageous decks in the Iowa Gambling Task (see Figure 2d). These findings are consistent with previous RNN modeling results and support the use of learned RNNs as black-box behavioral references for human trajectories (Kuperwajs et al., 2023; Ji-An et al., 2025; Eckstein et al., 2026; Yan et al., 2026). Full predictive metrics, including results for continuous decisions, are reported in Appendix B.

Behavioral alignment. We now turn to our primary evaluation objective: quantifying the human-likeness of behavioral trajectories generated by LLM agents using a learned human-policy RNN. For

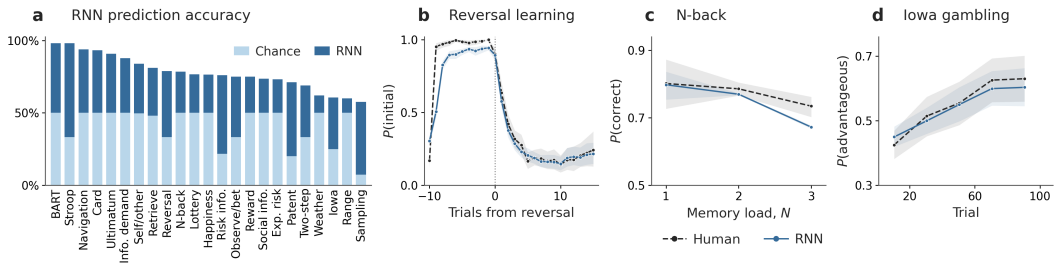

Figure 2: **Predictive validation of the human-policy RNN models. (a)** Held-out response prediction accuracy across the 23 tasks with categorical decisions. Full bar heights show RNN accuracy; pale blue marks uniform-choice chance levels. **(b–d)** Human responses and RNN probabilities for choosing the pre-reversal best option, responding correctly in N-back, and choosing advantageous Iowa decks, respectively. Predictions condition on recorded human histories. Shading denotes pointwise 95% participant-session cluster bootstrap intervals.

task ξ, we use $\pi_{\text{RNN}}(a_t \mid x_{1:t})$ to score the agent's action $a_t = \text{parse}_\xi(u_t)$, where $u_t \sim \pi_{\text{LLM}}(\cdot \mid c_t)$ and $x_{1:t}$ contains the current task inputs and the agent's own preceding interaction history. The RNN thus serves as a learned proxy for how human participants might respond under agent-generated histories, including those rarely or never observed in the human data. This evaluation relies on the RNN's ability to generalize within the task beyond the human trajectories on which it was trained.

Specifically, let $\tau^A = (o_1^A, a_1^A, o_2^A, a_2^A, \ldots, o_T^A, a_T^A)$ denote an LLM agent's trajectory, consisting of T observations and corresponding actions. At each decision step, we evaluate the agent's action using $\pi_{\text{RNN}}(a_t^A \mid x_{1:t}^A) = \pi_{\text{RNN}}(a_t^A \mid h_t^A)$, where h_t^A is the RNN's hidden state after processing the structured interaction history $(o_1^A, a_1^A, \ldots, o_t^A)$, before observing action a_t^A. We calculate a human-likeness score based on mean log likelihoods, which measure the likelihood assigned to the agent's actions by the RNN. Formally, the score $\bar{S} = \frac{1}{T}\sum_{t=1}^{T} \log \pi_{\text{RNN}}(a_t^A \mid h_t^A)$, which is the mean of the per-decision log likelihoods assigned by the RNN along the agent's trajectory (see Appendix E).

Optimal behavior in psychological tasks. Of the 25 psychological tasks, 17 have a clear definition of optimal behavior, including the Stroop and N-back tasks. In these tasks, behavior can be evaluated against an optimal decision a_t^* as a measure of task performance. The remaining tasks elicit preferences from human participants and agents, without a clearly defined optimality baseline. These 17 tasks therefore allow us to examine alignment with both human behavior and optimal behavior.

4 Experiments

We evaluated `mini-psych-agent` on 25 psychological tasks across GPT-5.6 Luna (OpenAI, 2026), DeepSeek V4.1 Flash (DeepSeek-AI, 2026), and Qwen3.8 Flash (Alibaba Cloud, 2026). Each backbone LLM was evaluated under five memory configurations (Stateless, Cumulative, Window, Retrieval, and Surprise), with five environment seeds per task, matched across configurations. Our analyses included 1,690 episodes across the 23 tasks with categorical decisions. Results for the 2 tasks with continuous or mixed decisions are reported in Appendix I.

For the subset of 17 tasks with a defined optimality baseline, we measured task performance as the proportion of decisions that exactly matched the task-specific optimal action. We assessed human-likeness using $\bar{S}$ and its random-to-human completeness, defined as $\frac{\bar{S}_{\text{agent}} - \bar{S}_{\text{random}}}{\bar{S}_{\text{human}} - \bar{S}_{\text{random}}}$. Comparisons between memory configurations were paired by seed within each task and backbone, and cross-task summaries weighted tasks equally. Unless otherwise stated, intervals are 95% confidence intervals obtained by bootstrapping tasks. Details of decoding and episode accounting are provided in Appendix C.

Memory effects on task performance. We evaluated how memory influences LLM agents' task performance, averaging task-specific optimality rates for cross-task aggregation. Across the 17 tasks, averaging the four memory configurations yielded a performance gain of 0.075 over Stateless (CI $[0.030, 0.123]$), with positive gains in 14 tasks. At the task–backbone level, individual memory

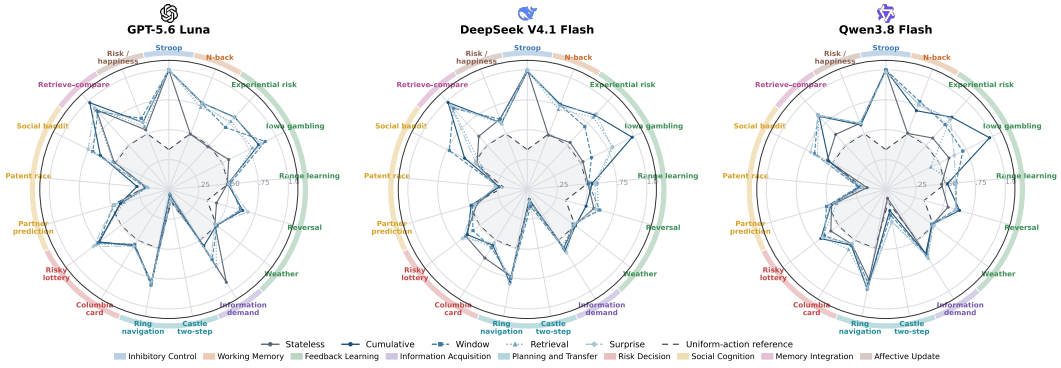

Figure 3: Jagged intelligence visualized using optimality rates across memory configurations and backbone LLMs on the 17 psychological tasks with clearly defined optimality baselines. Outer colors denote categories for psychological tasks, and lines represent memory configurations. Each axis shows a task-specific optimality rate, with 1 indicating that every decision matches an optimal decision. For the Risky lottery and Risk / happiness tasks, optimality is defined by the risk-neutral choice with the higher expected value; ties and decisions with ambiguous probabilities are excluded. The dashed boundary and gray region show the optimality baseline for an agent that selects uniformly among the available actions.

configurations improved performance relative to Stateless in 149 of 204 comparisons (73.0%), reduced it in 45, and tied in 10. Qwen showed consistent gains, improving in 15 of 17 tasks, with a mean gain of 0.084 (CI $[0.048, 0.123]$).

As shown in Figure 3, more specific patterns emerged across tasks with different history requirements. In experiential risk learning, Cumulative outperformed Window in all 15 paired backbone–seed comparisons, with a mean gain of 10.9 percentage points in optimal-choice rate. In Castle two-step, its optimal-terminal rate was 1.5 percentage points lower than Window's on average, with positive differences in 4 of 15 comparisons. In N-back, Window outperformed Retrieval by 3.3 percentage points in balanced accuracy, with gains in all 15 comparisons. These contrasts suggest benefits from accumulated experience in experiential risk learning and from preserving recent observations in their original order in N-back. The best memory configuration nevertheless varied across tasks and backbone LLMs (Appendix F).

Memory effects on human-likeness. We find that including interaction history in the context supplied to the LLM improves average alignment with the learned human policy. Across the 23 tasks, mean completeness increased from 0.720 under Stateless to 0.938 when averaged over the four other memory configurations, which provide interaction history in different forms ($\Delta = 0.218$, CI $[0.116, 0.334]$). Mean log likelihood $\bar{S}$ also increased by 0.167 nats per decision (CI $[0.094, 0.246]$). All four memory configurations had higher task-averaged completeness than Stateless for every backbone, with Cumulative scoring highest (Table 1). Averaged across the four configurations, completeness gains were 0.111 for GPT, 0.286 for DeepSeek, and 0.258 for Qwen. Holm-corrected Wilcoxon tests yielded $p = 0.063$, $p = 0.012$, and $p < 0.001$, respectively (Appendix D).

Task-specific memory effects were also evident as measured by $\bar{S}$. Cumulative exceeded Window by 0.234 nats in experiential risk learning and 0.251 nats in Castle two-step. In N-back, Window exceeded Retrieval by 0.127 nats, with positive differences in all 15 paired backbone–seed comparisons. These patterns suggest that the benefits of memory depend on the history demands of particular tasks, although better human-policy fit need not coincide with better task performance, as illustrated by Castle two-step.

Task-level completeness profiles, which express gains in human-policy fit over the random baseline relative to the held-out human reference, are shown in Figure 4. Compact memory configurations captured much of the aggregate benefit. Window and Surprise retained 72% and 85% of Cumulative's mean completeness gain over Stateless while using 9% and 34% of its prompt tokens, respectively, including memory-update calls. These ratios are computed from task-averaged gains and token use (see Appendix C for details). Cumulative's mean completeness marginally exceeded Win-

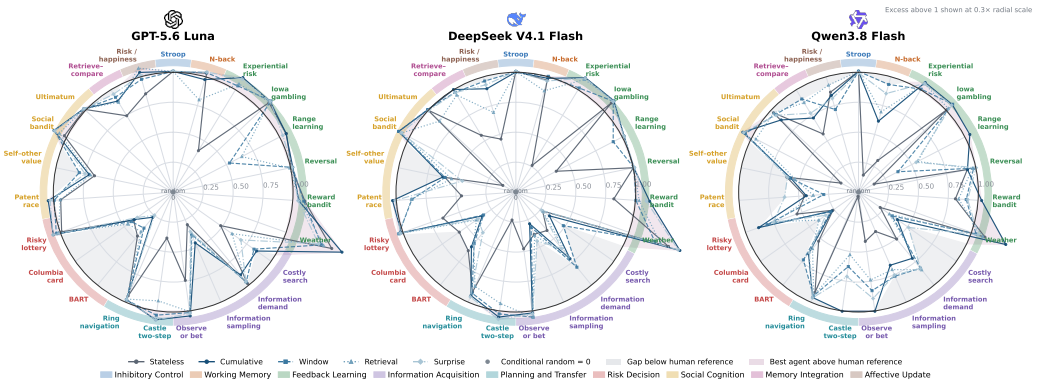

Figure 4: Jagged human alignment visualized using random-to-human completeness across the 23 psychological tasks with categorical decisions. The random reference averages RNN log likelihoods uniformly over valid actions along agent-visited histories; the human reference scores held-out participants on their own histories. Completeness of 0 therefore corresponds to the conditional random baseline, and 1 to the average held-out human log likelihood under the task-specific RNN. The human reference is a benchmark rather than an upper bound, and values are not clipped. Gray shading shows the shortfall of the highest agent completeness relative to 1; purple shading indicates completeness above 1. Excess above 1 is displayed with a radial scaling factor of 0.3. Outer colors denote psychological task categories, and lines represent memory configurations.

Table 1: Average random-to-human completeness across the 23 psychological tasks with categorical decisions. Brackets show 95% confidence intervals obtained by bootstrapping tasks; boldface indicates the highest mean for each backbone LLM. Within each task, completeness is computed from episode-averaged log likelihoods using matched seeds, with 1 indicating that the agent's $\bar{S}$ matches the held-out human reference score.

Model	Stateless	Cumulative	Window	Retrieval	Surprise
GPT-5.6 Luna	0.904 [0.740, 1.092]	**1.071** [0.901, 1.264]	0.999 [0.855, 1.149]	0.964 [0.824, 1.116]	1.024 [0.877, 1.169]
DeepSeek V4.1 Flash	0.682 [0.519, 0.881]	**1.010** [0.835, 1.201]	0.945 [0.812, 1.079]	0.928 [0.783, 1.080]	0.987 [0.812, 1.178]
Qwen3.8 Flash	0.574 [0.432, 0.723]	**0.897** [0.757, 1.051]	0.801 [0.690, 0.911]	0.789 [0.687, 0.895]	0.840 [0.745, 0.937]

dow's by 0.077, although the paired task-level difference was not statistically significant ($p = 0.273$; see Appendix D for details).

Relationship between task performance and human-likeness. Task performance and completeness were positively associated on average. Within each task and backbone, we computed Spearman's correlation between mean performance and mean completeness across all five memory configurations, then averaged these correlations across backbones within each task. Across the 17 tasks with both measures, Spearman's ρ was 0.412 ± 0.348 (mean $\pm$ SD across tasks; CI $[0.251, 0.573]$).

However, not all tasks showed positive correlation between task performance and human-likeness from agent's trajectories. For instance, N-back illustrates how the two measures can diverge within a sequential task. GPT's balanced accuracy increased from $(49.7 \pm 0.4)\%$ under Stateless to $(77.5 \pm 1.6)\%$ under Cumulative (mean $\pm$ SD across five seeds), while $\bar{S}$ decreased from -0.365 to -0.472 nats and completeness fell from 1.101 to 0.975. Specifically, Stateless responded with "non-match" on 96.3% of trials. An always "non-match" policy likewise achieved chance-level balanced accuracy but a score of $\bar{S} = -0.249$, exceeding the held-out humans' mean score of $\bar{S} = -0.458$. These results illustrate how repetitive choices can receive high likelihood under a history-conditioned human-policy model. Completeness therefore expresses mean log likelihood relative to the random and human references; it complements task performance but does not measure a distance between behavioral distributions (see Appendix B).

Interaction histories and autonomous behavior. We compared replay and closed-loop interaction across four tasks, two backbones (DeepSeek and Qwen), four memory configurations, and five seeds, yielding 160 matched episode pairs. In the replay condition, the environment followed a fixed trajectory sampled from the frozen human-policy RNN, and the forced actions were recorded in memory. Agent proposals were scored but did not alter the trajectory. Mean within-episode disagreement between proposed and forced actions ranged from 43.1% to 53.0% across configurations. In Castle two-step, the available second-stage options differed between matched closed-loop and replay trajectories in 66.6% of rounds. Mean memory gains in $\bar{S}$ were 0.387 and 0.267 nats under replay and closed-loop interaction, respectively (difference 0.120 nats; task-cluster bootstrap CI $[-0.121, 0.517]$). The highest-scoring configuration differed between conditions in three of the eight task–backbone combinations. These comparisons reveal how agents' own choices alter the states they encounter, alongside changes in memory configuration's rankings (see Appendix H).

We also compared complete trajectories from LLM agents with those generated autonomously by the frozen human-policy models across nine branching tasks. For each configuration, we generated 10,000 reference experiments, each matched to its initial scenarios and episode count. In Castle two-step, GPT with Cumulative memory achieved completeness of 1.249 but repeated its first-door choice on 95.8% of training transitions, compared with a reference mean of 49.5% (95% predictive interval $[39.0\%, 62.6\%]$). These checks reveal differences in sustained behavioral patterns even when the completeness measure is high.

5 Discussion

We introduced `MindGym`, a suite of psychology-inspired interactive environments for evaluating LLM agents. Compared with approaches that turn human experiences in psychological tasks into static prompts, `MindGym` supports closed-loop evaluation, where agents make their own decisions and influence subsequent task states. Agents can therefore be evaluated under their own interaction histories without repeatedly returning to a human participant's experience. This introduces an evaluation challenge: little or no human data may be available for states rarely or never visited by human participants. We address this by using task-specific RNNs trained on human trajectories as learned behavioral references for evaluating agent-generated trajectories.

Using `mini-psych-agent`, we systematically varied how LLM agents manage their external memory files and incorporate past interactions into the context at each decision step. Overall, retaining more past interactions in context generally improved both task performance and human-likeness, although no single memory strategy performed best across all tasks and backbones. Cumulative outperformed Window in feedback learning and planning tasks, while Window outperformed Retrieval in working memory tasks. These patterns suggest that memory benefits depend on both the amount of past experience retained and the preservation of recent temporal order. Window and Surprise nevertheless captured much of Cumulative's benefit for human-likeness while using substantially fewer tokens in prompts for decision making.

Alignment with optimal behavior and alignment with human behavior were found to be positively related, but could also diverge in certain tasks. This suggests that greater optimality can coincide with more human-like behavior in some cognitive domains. Agents with high human-likeness scores could nevertheless exhibit qualitatively distinct behavioral patterns over complete trajectories. Therefore, task performance, conditional human-policy fit, and behavior across complete trajectories provide complementary perspectives on closed-loop interaction, supporting a more holistic evaluation of autonomous agents.

Limitations and future research. Our environments and agents represent early steps toward psychology-inspired evaluation of LLM agents. Additional psychological tasks could be implemented as interactive environments within `MindGym` to examine a broader range of cognitive processes. More sophisticated memory strategies, particularly those that modify stored representations of individual trials, could be incorporated into `mini-psych-agent`, alongside additional skills and tool-use capabilities. Another promising direction is to train LLM agents to behave more like humans within psychological environments, using learned human-policy models as training targets. We believe these efforts could help develop autonomous agents whose decisions are better aligned with human behavior.

AI USE STATEMENT

We used generative AI tools to assist with implementing and debugging code. LLM-generated trajectories are experimental outputs of the evaluated agents and are distinguished from the human behavioral data used to train the RNNs. The authors take responsibility for the final methods, analyses, claims, and artifacts presented in this work.

ETHICS STATEMENT

This work uses previously collected human behavioral datasets and computational experiments with LLM agents. The learned human-policy models reflect patterns in their source populations and should not be interpreted as universal models of human cognition. The reported alignment scores are not intended for clinical assessment or individual psychological profiling, and higher scores do not necessarily imply human-like reasoning or optimal behavior. Data access and any redistribution remain subject to the original datasets' licensing, access, and privacy requirements.

REPRODUCIBILITY STATEMENT

Task protocols and environment adaptations are documented in Appendix A. RNN input encodings, data splits, training configurations, and predictive validation are described in Appendix B. Agent configurations, decoding settings, and episode accounting are reported in Appendix C. Score definitions and their interpretation are provided in Appendix E, while aggregation, seed matching, and statistical procedures are detailed in Appendix D.

We provide the code and experimental data for anonymous review.[1]

REFERENCES

Mayank Agrawal, Joshua C Peterson, and Thomas L Griffiths. Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences*, 117(16):8825–8835, 2020.

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390, 2025. doi: 10.1038/s41562-025-02172-y.

Alibaba Cloud. Qwen3.8-Flash. Alibaba Cloud Model Studio documentation, 2026. URL https://www.alibabacloud.com/help/en/model-studio/qwen3-8-flash. Accessed September 23, 2026.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122, 2025.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025.

Marcel Binz, Elif Akata, Abdullah Almaatouq, Mohammed Alsobay, Oleksii Ariasov, Franziska Brändle, David Broska, Jason W Burton, Nuno Busch, Frederick Callaway, et al. Post-training makes large language models less human-like. *arXiv preprint arXiv:2605.07632*, 2026.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016. URL https://arxiv.org/abs/1606.01540.

[1] https://anonymous.4open.science/r/mindgym-anonymous-review-FB11/

Yifei Cao and Maria K. Eckstein. Hybrid neural-cognitive models reveal flexible context-dependent information processing in reversal learning. *bioRxiv preprint*, 2025. doi: 10.1101/2025.07.31.667923.

Jiayang Cheng, Dongyu Ru, Lin Qiu, Yiyang Li, Xuezhi Cao, Yangqiu Song, and Xunliang Cai. AMemGym: Interactive memory benchmarking for assistants in long-horizon conversations. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=sfrVLzsmlf.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, 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 (EMNLP)*, pp. 1724–1734, 2014.

Julian Coda-Forno, Marcel Binz, Jane X. Wang, and Eric Schulz. CogBench: a large language model walks into a psychology lab. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9076–9108. PMLR, 2024. URL https://proceedings.mlr.press/v235/coda-forno24a.html.

Nathaniel D Daw. Trial-by-trial data analysis using computational models. *Decision making, affect, and learning: Attention and performance XXIII*, 23(1):3–38, 2011.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash/blob/main/DeepSeek_V41_Tech_Report.pdf.

Amir Dezfouli, Kristi Griffiths, Fabio Ramos, Peter Dayan, and Bernard W. Balleine. Models that learn how humans learn: The case of decision-making and its disorders. *PLOS Computational Biology*, 15(6):e1006903, 2019. doi: 10.1371/journal.pcbi.1006903.

Maria K. Eckstein, Christopher Summerfield, Nathaniel D. Daw, and Kevin J. Miller. Hybrid neural–cognitive models reveal how memory shapes human reward learning. *Nature Human Behaviour*, 10(5):972–987, 2026. doi: 10.1038/s41562-025-02324-0.

Kianté Fernandez, Caitlin Chen, Grace Zhou, Anthony Miceli, Bartek Sadowski, and Ian Krajbich. Cogarena: Benchmarking ai agents on interactive behavioral experiments. 2025. URL https://cog-arena.vercel.app.

Zafeirios Fountas, Martin Benfeghoul, Adnan Oomerjee, Fenia Christopoulou, Gerasimos Lampouras, Haitham Bou Ammar, and Jun Wang. Human-inspired episodic memory for infinite context LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BI2int5SAC.

Michael C Frank and Noah D Goodman. Cognitive modeling using artificial intelligence. *Annual Review of Psychology*, 77(1):543–566, 2026.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Leon Guertler, Bobby Cheng, Simon Yu, Bo Liu, Leshem Choshen, and Cheston Tan. TextArena. *arXiv preprint arXiv:2504.11442*, 2025. URL https://arxiv.org/abs/2504.11442.

Li Ji-An, Marcus K. Benna, and Marcelo G. Mattar. Discovering cognitive strategies with tiny recurrent neural networks. *Nature*, 644(8078):993–1001, 2025. doi: 10.1038/s41586-025-09142-4.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *arXiv preprint arXiv:2407.01502*, 2024. URL https://arxiv.org/abs/2407.01502.

Akshay Krishnamurthy, Keegan Harris, Dylan J. Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context? In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2403.15371.

Ionatan Kuperwajs, Heiko H Schütt, and Wei Ji Ma. Using deep neural networks as a guide for modeling human planning. *Scientific reports*, 13(1):20269, 2023.

Yuxuan Lei, Jianxun Lian, Defu Lian, Jincenzi Wu, Tianfu Wang, and Xing Xie. A computational framework for evaluating human-likeness in llms' open-ended human behaviors. *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Joel Z. Leibo, Cyprien de Masson d'Autume, Daniel Zoran, David Amos, Charles Beattie, Keith Anderson, Antonio García Castañeda, Manuel Sanchez, Simon Green, Audrunas Gruslys, Shane Legg, Demis Hassabis, and Matthew M. Botvinick. Psychlab: A psychology laboratory for deep reinforcement learning agents. *arXiv preprint arXiv:1801.08116*, 2018. URL https://arxiv.org/abs/1801.08116.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2005.11401.

Ryan Liu, Jiayi Geng, Joshua Peterson, Ilia Sucholutsky, and Thomas L. Griffiths. Large language models assume people are more rational than we really are. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 49900–49926, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/7cdb411d07bdf6cd43ddf1254fae98cd-Paper-Conference.pdf.

George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

Manuel Molano-Mazón, Joao Barbosa, Jordi Pastor-Ciurana, Marta Fradera, Ru-Yuan Zhang, Jeremy Forest, Jorge del Pozo, Li Ji-An, Christopher J. Cueva, Jaime de la Rocha, Devika Narain, and Guangyu Robert Yang. NeuroGym: An open resource for developing and sharing neuroscience tasks. *PsyArXiv*, 2022. doi: 10.31234/osf.io/aqc9n. URL https://doi.org/10.31234/osf.io/aqc9n.

OpenAI. GPT-5.6 Luna. OpenAI API documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.6-luna. Accessed September 23, 2026.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. URL https://arxiv.org/abs/2310.08560.

Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, Jakob Nicolaus Foerster, Jack Parker-Holder, and Tim Rocktäschel. BALROG: Benchmarking agentic LLM and VLM reasoning on games. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.13543.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *ACM Symposium on User Interface Software and Technology*, 2023. URL https://arxiv.org/abs/2304.03442.

Weihan Peng, Chenxu Zhang, Qianao Wang, Yuling Shi, Heng Lian, Qihong Mao, Jiahao Pang, Chunliang Feng, Bowen Li, and Xiaodong Gu. Heart-bench: Do llm agents exhibit human-like psychology? *arXiv preprint arXiv:2605.30058*, 2026.

Joshua C. Peterson, David D. Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L. Griffiths. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214, 2021. doi: 10.1126/science.abe2629.

Johannes A. Schubert, Akshay K. Jagadish, Marcel Binz, and Eric Schulz. In-context learning agents are asymmetric belief updaters. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43928–43946. PMLR, 2024. URL https://proceedings.mlr.press/v235/schubert24a.html.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2309.02427.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024. URL https://arxiv.org/abs/2407.17032.

Endel Tulving and Donald M Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological review*, 80(5):352, 1973.

Robert C Wilson and Anne GE Collins. Ten simple rules for the computational modeling of behavioral data. *elife*, 8:e49547, 2019.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.10813.

Hanbo Xie and Jian-Qiao Zhu. Centaur may have learned a shortcut that explains away psychological tasks. *PsyArXiv*, 2025.

Haijiang Yan, Jian-Qiao Zhu, Liqiang Huang, and Ming Meng. Understanding human-like solutions in combinatorial optimization via learning and search. *arXiv preprint arXiv:2607.23854*, 2026.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.15793.

Lance Ying, Jinzhou Wu, Yingshan Susan Wang, Shivam Aarya, Luca M. Schulze Buschoff, Harry Chen, Katherine M. Collins, Andrea de Varda, Shuhao Fu, Sean Dae Houlihan, et al. CogGym: Towards large-scale comparative evaluation of human and machine cognition. *arXiv preprint arXiv:2609.21259*, 2026. URL https://arxiv.org/abs/2609.21259.

Jian-Qiao Zhu and Thomas L Griffiths. Incoherent probability judgments in large language models. *Proceedings of the 46th Annual Meeting of the Cognitive Science Society*, 2024.

Jian-Qiao Zhu, Joshua C. Peterson, Benjamin Enke, and Thomas L. Griffiths. Capturing the complexity of human strategic decision-making with machine learning. *Nature Human Behaviour*, 9 (10):2114–2120, 2025. doi: 10.1038/s41562-025-02230-5.

Brian D Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University, 2010.

Sam Zorowitz and NIMH DSST. OpenCogData. Online database of cognitive task datasets, 2024. URL https://nimh-dsst.github.io/OpenCogData/. Accessed: 2026-09-26.

A ILLUSTRATIONS FOR INTERACTIVE PSYCHOLOGICAL ENVIRONMENTS IN MINDGYM

Each environment exposes task instructions, a textual observation, and the currently legal responses. Executing a response updates the task state and returns the next observation, any protocol-permitted feedback, and a termination signal. The observation contains participant-visible information: for example, previously revealed outcomes, current offers, or a revealed portion of a payoff matrix. Hidden reward parameters, unrevealed outcomes, and evaluator-only optimality calculations are kept separate. The number of environment steps need not equal the number of experimental trials: information search, navigation, and multistage decisions can require several actions within one trial or round.

Tables 2–3 summarize the 25 implemented tasks. All are untimed text adaptations; visual displays are described symbolically and response times are not modeled. Source schedules and feedback conditions are retained where available, while interface instructions and cover stories are not uniformly reproduced. Iowa gambling, momentary mood, and sparse-event estimation include declared reconstructions where original implementation details are unavailable.

B HUMAN-POLICY MODELS AND PREDICTIVE VALIDATION

Each task has a separately fitted RNN model on training human trajectories. Encodings represent visible stimuli, protocol stage, and causally available response/outcome history (Table 4). Discrete identities use indicators; numerical features use task-specific scaling. Candidate-based choices pool masked candidate encodings before recurrent updating, with shared utility heads scoring legal options. Previous responses are shifted: the probability of response t is computed before that response enters history. Observation-only events can update recurrent state without contributing a response loss. State resets at source sequence boundaries.

Categorical heads minimize binary or multiclass cross-entropy, masking illegal actions and unscored events. Retrieve-compare uses separate two-choice and 21-category report heads; costly search uses conditional typed heads. Mood uses mean squared error; sparse-event estimation uses squared error normalized by protocol slider range. Happiness and self–other models combine categorical losses with a weight-10 smooth-L_1 rating loss on the $[0, 1]$ scale. Continuous outputs are evaluated in native units. Adam-family optimization and validation-based checkpoint selection are used; the table records selected widths, learning rates, batch sizes, maximum epoch budgets, and stopping patience. Maximum epochs are budgets, not fitted epoch counts. Held-out groups are source-defined participant/session identities in the saved splits.

Across 23 categorical tasks, mean saved test NLL is 0.504 ± 0.299 nats (task mean $\pm$ task standard deviation). The participant-sequence-balanced human anchor, averaged equally across tasks, is $\bar{S} = -0.554$: it uses sequence means and agent-scored endpoints, including joint rather than head-averaged likelihood for costly search. Thus the summaries have different aggregation and endpoint conventions. Table 5 reports categorical and continuous test metrics. Categorical trajectory scores use a probability floor of 10^{-12}.

A sampled-reference check covers 16 tasks: the mean absolute task-level difference from the held-out human participants $\bar{S}$ is 0.030 nats, with 13 tasks below 0.05. Fourteen tasks average five environment seeds; N-back and Stroop average 64 and four unique yoked stimulus schedules, respectively. Sampled responses condition on their own preceding responses; yoked schedules are retained rather than regenerated adaptively.

N-back illustrates why predictive likelihood complements performance. Always choosing nonmatch yields $\bar{S} = -0.249$ but balanced accuracy 0.500, compared with the human anchor $\bar{S} = -0.458$. GPT Stateless chooses nonmatch on 96.3% of responses; cumulative memory raises balanced accuracy from 49.7% to 77.5% while lowering $\bar{S}$ from -0.365 to -0.472. Predictable repetition can receive high conditional likelihood without reproducing the human behavioral distribution.

Table 2: Environment inventory (1/2). Counts distinguish choices, trials, and rounds; a round may contain multiple actions.

Task	Interaction and task-specific adaptation
Stroop	Report the annotated color of a color word from three alternatives over 84 trials. Ink color is described in text, so the interface does not reproduce visual interference or response-time effects.
N-back	Judge letter matches over 20 adaptive blocks, each with 20 scored positions and N initial context positions. Load changes between blocks; trial feedback and the source 0-back control are absent.
Experiential risk	Choose between labeled stimuli in 110 free choices, interleaved with 66 forced-exposure events presented as feedback. Some outcome pools deplete with use; stimulus images and spatial layout are abstracted.
Iowa gambling	Make 100 draws from four decks, observing gains, losses, and balance. Published payoff tables instantiate source-specific schemes; finite-deck depletion includes an explicitly reconstructed rule.
Range learning	Choose between labeled symbols over 120 learning and 120 transfer trials. Source conditions determine block structure and partial or complete outcome feedback; practice is omitted.
Reversal learning	Choose among three machines over four 40-choice blocks. Only the selected outcome is shown; recent choices can trigger reward-probability reversals. Machines replace the source card/deck display.
Reward bandit	Make 100 binary choices under hidden fixed reward schedules. The agent sees only the chosen reward outcome, remaining choices, and accumulated rewards; graphical rewards are rendered as text.
Weather	Predict between two alternatives from cue counts replacing icons. Source sessions determine the cue sequence and length; feedback stages reveal the rewarded side, while test stages withhold outcome feedback.
Information sampling	Reveal boxes on a 5×5 text board or choose its majority color. There are 10 fixed-reward and 10 decreasing-reward rounds; opening boxes reduces correct-answer rewards only in the latter condition.
Information demand	Four 126-trial blocks combine lottery queries with Observe, Estimate, or Intervene responses. Bonuses are settled at block end; the middle-block order is seeded, and source practice/quiz stages are omitted.
Observe-or-bet	Choose to observe a light or bet on red/blue: 25 practice choices followed by two 50-choice paying blocks. Bets hide immediate outcomes; observations reveal color and block ends reveal total score.
Costly search	Across 20 rounds, pay to reveal cells of a four-color, six-gamble payoff matrix, then choose a gamble. Text replaces the query interface; color probabilities and query charges are public.
Castle two-step	Learn stable door mappings across 100 training rounds and one transfer round. Each round has two three-door choices; the terminal resource bundle and price-weighted reward appear after the second choice.

Table 3: Environment inventory (2/2). Subjective reports and preference choices remain distinct from objective task-performance measures.

Task	Interaction and task-specific adaptation
Ring navigation	Move clockwise/counterclockwise on a text-described 9- or 10-node ring. Evaluated sessions contain 80 goal-reaching training rounds and one test probe, which terminates after its first valid move.
Risky lottery	Choose between two lotteries for 20 problems repeated five times. Payoffs, disclosed probabilities, and source feedback conditions are rendered as text; some problems contain undisclosed probabilities.
BART	Pump or cash out across 30 balloons. Each pump adds a pending point, while explosion removes that balloon's pending points; hidden source explosion thresholds are randomly ordered.
Columbia card	Turn another card or stop over 84 rounds, initially containing 32 cards each. Gain, loss, and loss-card count are public; sequential card outcomes and terminal scores are presented as text.
Patent race	Invest 0–4 from a four-unit endowment over 80 rounds. A strictly larger investment wins a ten-unit prize; opponent investments are sampled from round-specific recorded-choice pools and revealed afterward.
Self-other	Make 18 own-allocation choices, 36 partner-choice predictions, two intention ratings, and one attribution. Allocations are text; recorded partner choices and prediction feedback provide the social-learning sequence.
Social bandit	Choose a lake over three 15-choice blocks with private outcomes and condition-specific observation/advice. Social information appears at its scheduled event; the source's additional final stage is omitted.
Ultimatum	Accept or reject 20 shuffled offers dividing 100 points. Acceptance implements the offer; rejection yields zero to both players. Only responder choices are retained, excluding source emotion ratings and third-party roles.
Retrieve-compare	Learn labeled-symbol values, then make transfer choices and probability reports. Source-scheduled practice and feedback conditions are retained; pie charts become stated probabilities and reports use 21 levels from 0% to 100%.
Risk/happiness	Make 30 safe-versus-risky choices interleaved with 12 happiness reports, beginning before the first choice. Offers and outcomes are text; happiness uses a 0–100 response scale.
Momentary mood	Report baseline and post-reward mood on a 101-level scale from -1 to 1. Mood-dependent rewards and progression use a declared reconstruction; screening and stopping can make session length variable.
Sparse estimation	Record numerical events in two 40-event contexts, then estimate both means. The untimed reconstruction includes source-dependent fillers/probes; integer-entry scales are public, and the human reference scores the final estimates.

C Agent Configurations and Episode Accounting

Memory and decoding. Window and Retrieval provide at most six interaction records; Retrieval limits each retrieved interaction to 2,000 characters. Surprise uses the preceding ten valid forecast scores for the same target, with a five-score warm-up, and closes an event when surprise exceeds their mean plus one standard deviation. Events contain at least two and at most 32 completed decisions; summaries contain at most 1,800 characters. The current event remains uncompressed. All three backbones run without extended reasoning. DeepSeek and Qwen use temperature 1 and top-p 1; GPT uses provider-default sampling. Actions are returned as structured JSON and checked against the current legal action set. Invalid or truncated actions terminate the episode without executing a substitute action. Failed runs are recorded separately; retained episodes contain no fallback actions.

Episode accounting. The design specifies 1,875 episodes: 25 tasks, three backbones, five configurations, and five environment seeds. Of these, 1,812 completed, 50 terminated at the momentary-mood responsiveness screen, and 13 were incomplete. The latter comprise six sparse-estimation episodes and seven categorical-task episodes. Among the 1,718 completed categorical episodes, 1,690 have a matching seed in every configuration and enter the 23-task human-policy analysis. The full matched performance cohort contains 1,735 episodes; its 19 tasks with verified random and ideal references contribute 1,360 episodes to normalized comparisons. Environment seeds match initial conditions across configurations; model sampling and subsequent trajectories can differ.

Context usage. Table 6 reports prompt tokens per environment decision in the matched categorical cohort. Each episode's total prompt tokens, including forecast and summary calls, is divided by its decision count; these ratios are averaged over seeds and then equally over tasks. Pooled comparisons first average backbones within each task. Token shares are ratios of these means, and retained completeness gains are ratios of mean improvements over Stateless. Window uses 8.72% of Cumulative's prompt tokens and retains 71.65% of its completeness gain; Surprise uses 34.37% and retains 84.67%. These are context-use measures, not monetary costs or cache-adjusted billing estimates.

D Aggregation and Exploratory Inference

Analyses retain complete, valid episodes from seeds available under all five configurations within each task and backbone. This yields 1,690 episodes for categorical human-policy scoring and 1,360 episodes for the 19-task normalized-performance analysis. Matching seeds controls initial scenarios, not the subsequent action-dependent trajectories.

Performance is normalized within episode as $(v - v_{\mathrm{random}})/(v_{\mathrm{ideal}} - v_{\mathrm{random}})$, using the corresponding task references, and then averaged over matched seeds. For completeness, we first average episode-level agent and conditional-uniform log scores, then compute completeness as $(\overline{S}_{\mathrm{agent}} - \overline{S}_{\mathrm{random}})/(\overline{S}_{\mathrm{human}} - \overline{S}_{\mathrm{random}})$. Thus, the reported cell score is a ratio of means, not a mean of episode ratios. The human anchor is the equal-sequence mean of held-out human scores. Neither score is clipped.

"Any memory" averages the four memory configurations before subtracting Stateless. Pooled effects average available backbones within each task, then weight tasks equally. Confidence intervals are the 2.5th and 97.5th percentiles of 10,000 bootstrap means obtained by resampling tasks (random seed 20260925), retaining their configurations and backbones together. These intervals hold the fitted scorers and human anchors fixed and are not multiplicity-adjusted. The resulting pooled effects are 0.140 $[0.056, 0.234]$ for performance, 0.218 $[0.116, 0.334]$ for completeness, and 0.167 $[0.094, 0.246]$ nats for $\bar{S}$.

Completeness effects are tested using two-sided task-level Wilcoxon signed-rank tests after removing exact zero differences. The Holm family contains eight comparisons: any-memory effects for three backbones, four pooled memory-versus-Stateless effects, and the pooled any-memory effect. Backbone results are shown below. For the separate Cumulative–Window comparison, the main text reports the unadjusted $p = 0.273$; Holm adjustment over all six memory-pair contrasts for each of three backbones and the pooled analysis gives $p = 1.000$.

For the 18 categorical tasks with normalized performance, Spearman correlations are computed across configurations within each task–backbone pair, then averaged across available backbones and tasks. Undefined correlations from constant profiles are omitted. Task bootstrap gives $\bar{\rho} = 0.384$ $[0.205, 0.550]$; omitting Stateless gives 0.290 $[0.113, 0.467]$. All analyses are exploratory; individual decisions are not treated as independent replicates.

E USING LEARNED HUMAN-POLICY MODEL TO EVALUATE AGENT-GENERATED TRAJECTORIES

We evaluate an agent's decisions using the probabilities assigned by a learned human-policy model along the agent's own interaction history. Below, we derive this score from the trajectory likelihood and clarify its interpretation.

For a fixed task ξ, consider an agent-generated trajectory $\tau^A = (o_1^A, a_1^A, \ldots, o_T^A, a_T^A)$. Let $\eta_t^A = (o_1^A, a_1^A, \ldots, o_t^A)$ denote the public interaction history available before action a_t^A, and let $x_{1:t}^A$ denote its structured representation. The frozen RNN processes $x_{1:t}^A$ to obtain its hidden state h_t^A and assigns the observed action probability

$$\pi_{\text{RNN}}(a_t^A \mid x_{1:t}^A) = \pi_{\text{RNN}}(a_t^A \mid h_t^A). \tag{6}$$

The recurrent state is updated using the agent's preceding actions and feedback. The current action a_t^A enters the history only after its probability has been evaluated. Thus, the RNN scores decisions under histories generated by the agent, without requiring a matching human trajectory.

For categorical decisions and a fixed horizon T, the probability of the same trajectory under the RNN policy interacting with the environment factorizes as

$$P_{\text{RNN}}^{\xi}(\tau^A) = \underbrace{\mu_\xi(o_1^A) \cdot \prod_{t=1}^{T-1} P_\xi(o_{t+1}^A \mid \eta_t^A, a_t^A)}_{\text{environment contribution}} \cdot \underbrace{\prod_{t=1}^{T} \pi_{\text{RNN}}(a_t^A \mid h_t^A)}_{\text{action contribution}}, \tag{7}$$

where μ_ξ is the initial-observation distribution and P_ξ is the distribution of the next observation induced by the environment dynamics and rendering function. Conditioning P_ξ on the participant-visible history accommodates environments whose internal states are only partially observed.

Our evaluation uses the action contribution,

$$S_{\text{RNN}}(\tau^A) = \prod_{t=1}^{T} \pi_{\text{RNN}}(a_t^A \mid h_t^A). \tag{8}$$

This is a causal action likelihood: each action is evaluated using only information available before that decision (Ziebart, 2010). It is generally different from conditioning the entire action sequence on the entire observation sequence, which would also condition earlier decisions on future observations. The factorization does not assume independent actions, because each probability depends on the preceding interaction history.

Scoring the action contribution evaluates the agent's choices while excluding the direct likelihood of the environment's stochastic outcomes. Those outcomes still affect subsequent scores through the interaction history. The environment contribution can differ across trajectories, so S_{RNN} should not be interpreted as the full joint trajectory probability.

To express the score per decision, we take the logarithm and divide by the number of decisions:

$$\bar{S}_{\text{RNN}}(\tau^A) = \frac{1}{T} \log S_{\text{RNN}}(\tau^A) \tag{9}$$

$$= \frac{1}{T} \sum_{t=1}^{T} \log \pi_{\text{RNN}}(a_t^A \mid h_t^A). \tag{10}$$

Exponentiating gives the geometric mean of the per-decision action probabilities:

$$\exp\big(\bar{S}_{\text{RNN}}(\tau^A)\big) = \left[\prod_{t=1}^{T} \pi_{\text{RNN}}(a_t^A \mid h_t^A)\right]^{1/T}. \tag{11}$$

Thus, $\exp\left(\bar{S}_{\text{RNN}}\right)$ is the geometric mean of likelihoods. They produce identical rankings because exponentiation is strictly increasing. This construction preserves the multiplicative structure of sequential action likelihoods while expressing the result on a per-decision scale. For variable-length episodes, we use each episode's number of decisions in the denominator. This normalization removes the direct dependence of the product on sequence length, although differences in task difficulty and visited histories remain.

Interpreting the score. The logarithmic score is strictly proper when used to evaluate a predictive distribution against responses drawn from a fixed target distribution (Gneiting & Raftery, 2007). This supports its use for training and validating the RNN on human trajectories. Evaluating agent-generated trajectories under a fixed RNN, however, measures their compatibility with the learned human policy rather than guaranteeing distributional matching.

To make this distinction explicit, let $\pi_{\text{Agent}}(a_t \mid \eta_t)$ denote the agent's induced policy over executed actions, including the effects of its LLM and harness. For a fixed horizon T, define its average conditional action entropy as

$$\mathcal{H}_{\text{Agent}} = -\frac{1}{T}\mathbb{E}_{\tau\sim P^{\xi}_{\text{Agent}}}\left[\sum_{t=1}^{T}\log \pi_{\text{Agent}}(a_t \mid \eta_t)\right]. \tag{12}$$

When the agent and RNN policies interact with the same environment and initial-observation distribution,

$$\mathbb{E}_{\tau\sim P^{\xi}_{\text{Agent}}}\left[\bar{S}_{\text{RNN}}(\tau)\right] = -\mathcal{H}_{\text{Agent}} - \frac{1}{T}D_{\text{KL}}\left(P^{\xi}_{\text{Agent}} \| P^{\xi}_{\text{RNN}}\right). \tag{13}$$

This identity follows by expanding the trajectory KL divergence: the shared environment factors in Equation 7 cancel, leaving the sum of action log-probability ratios.

A higher expected score can therefore reflect both a smaller divergence from the reference policy and lower action entropy. In particular, an agent that consistently chooses modal human responses can score higher than one that reproduces human variability. We consequently interpret the score as behavioral compatibility with a learned human policy. Computing it requires only the agent's sampled actions and the RNN's probabilities, without access to the agent's own action probabilities. Obtaining the latter can be difficult even when token-level probabilities are available, because multiple response strings and reasoning traces can lead to the same parsed action. Computing an action's probability generally requires marginalizing over these alternatives; the probability of an answer token conditioned on a single reasoning trace need not equal the agent's marginal action probability. Moreover, interpreting the mean log likelihoods as human-likeness also depends on the RNN's predictive validity under agent-generated histories. Performance on held-out human trajectories supports the use of the RNN as a reference, but does not by itself establish reliable extrapolation to rarely or never observed histories.

F Exploratory statistical analysis of task performance

We analyzed the existing scalar performance endpoints of all 19 tasks with a defined improvement direction. The comparison family comprised 228 contrasts: four memory configurations versus Stateless for each of three backbones and 19 tasks. Within each task and backbone, we retained seeds with complete, finite outcomes under all five configurations and required at least three matched seeds, yielding 224 estimable contrasts. Each episode contributed one summary outcome; sequential decisions were not treated as independent replicates. Positive differences indicate improved performance, with signs reversed for lower-is-better endpoints.

We report mean paired differences, marginal unadjusted 95% paired-t confidence intervals, and both two-sided paired-t and exact sign-flip tests of the mean difference. Holm adjustment was applied separately to each test procedure across the same 228 planned contrasts. Undefined tests were flagged and assigned $p = 1$ solely for adjustment bookkeeping. Paired-t inference assumes independent, approximately normal seed differences; sign-flip inference requires the corresponding null symmetry or within-pair exchangeability. With five pairs, the usual two-sided exact test cannot yield $p < 0.0625$. Small sample sizes limit assessment of these assumptions. The confidence intervals are not multiplicity-adjusted.

These analyses were specified after inspection of descriptive outcomes and are exploratory. Adjustment across the full comparison family does not remove prior task or reporting selection. Matching protocol seeds does not imply identical realized trajectories in adaptive or action-dependent tasks. Uncertainty is conditional on the evaluated tasks, backbones, protocol, frozen endpoints, and retained complete runs; neither test is treated as independent confirmation or selected according to its significance.

G THE BEST MEMORY CONFIGURATION DEPENDS ON THE TASK

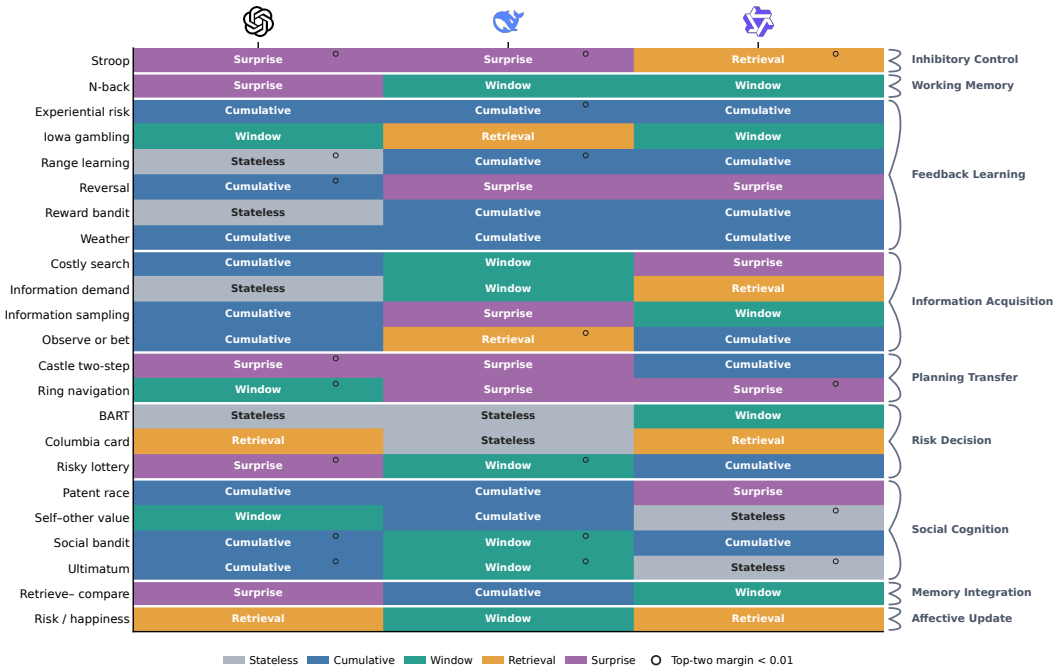

Figure 5: Task-level variation in human-policy likelihood using the latest audited result table. Each cell identifies the memory configuration with the highest $\bar{S}$ for a task and backbone. Open circles mark cells for which the difference between the two highest means is below 0.01 nats per scored action. Colors encode descriptive mean winners and do not imply statistically significant pairwise superiority.

No single memory mechanism is the winner for every task and backbone (Figure 5). Although Cumulative memory performs best on average, the configuration that most closely matches the learned human policy in raw likelihood is task- and backbone-dependent; many near-ties further caution against interpreting small mean differences as reliable superiority.

H REPLAY AND AUTONOMOUS BEHAVIOR

Teacher-forced replay We paired 160 episodes across four tasks, two backbones, four memory configurations, and five seeds. For each task and seed, all conditions shared one trajectory sampled from the frozen human-policy RNN. Replay executed its actions and recorded them in memory; agent proposals were scored without changing the path. Closed-loop interaction executed the agent's own actions. Table 8 reports $\bar{S}$ under each mode's realized history. Memory gain averaged Cumulative, Window, and Retrieval relative to Stateless. Resampling tasks and then paired seeds, keeping both backbones together, gave a replay-minus-closed-loop gain of 0.120 nats (10,000 bootstrap draws; 95% interval $[-0.121, 0.517]$).

Autonomous behavioral checks. We compared 623 included episodes across nine branching tasks with complete autonomous RNN trajectories. For each task–backbone–memory group, 10,000

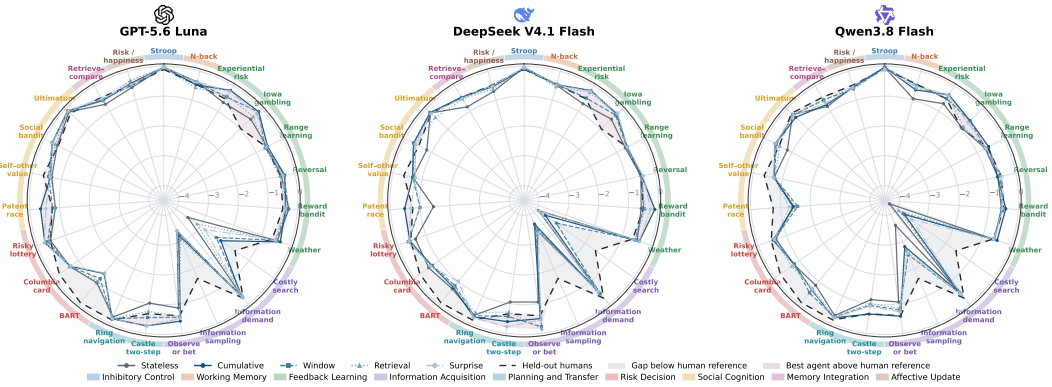

Figure 6: Human-policy fit $\bar{S}$ across the 23 categorical tasks. Each axis reports $\bar{S}$ in nats per scored action, so values closer to zero indicate that the learned human policy assigns greater likelihood to the agent's choices. The dashed line is the score of held-out human sequences and is an empirical reference rather than an upper bound. Outer colors denote psychological categories; lines denote memory configurations. Gray shading marks the gap between the best displayed agent configuration and the human reference; purple shading marks tasks on which the best agent score exceeds that reference. Because raw $\bar{S}$ depends on task-specific action spaces and predictability, radial distances and polygon areas should not be compared as aggregate scores across tasks.

reference experiments matched its initial scenarios and episode count, drawing one full trajectory per scenario. Ordinary features used equal-episode means; conditional rates pooled event and opportunity counts. Undefined statistics remained unavailable. Figure 7 shows the switching and repetition features discussed in the main text, retaining all configurations. Bands are predictive intervals under the frozen RNN, rather than confidence intervals for LLM effects or human population means. Complete feature results are provided in the supplementary data.

I Results on Continuous-Response Tasks

For Momentary mood and Sparse estimation, we report mean absolute deviation (MAD) between agent responses and the frozen human-policy model's predictions, conditioned on the agent's own interaction history. Lower MAD indicates closer agreement. Mood scores use the $[-1, 1]$ scale; Sparse-estimation MAD averages deviations for the post-training and post-test estimates in original slider points (ranges 0–300 or 0–500). Table 9 averages episode-level MAD over completed episodes. Momentary mood has 25 completed episodes and 50 protocol-screened terminations; Sparse estimation has 69 completed episodes and six incomplete episodes. Cell counts are reported because the available seeds differ across configurations.

J Prompt examples

The following abridged example shows the participant-facing text for the first Stroop decision in a saved Stateless episode (seed 101). Line breaks are adjusted for readability. The task instructions and current observation are supplied separately from any memory selected by the configuration; no previous interaction exists at this first step.

Task instructions (excerpt).

> This is a three-colour Stroop response task. On every trial, report the PRINTED COLOUR annotation, not the word's meaning. No response-time measure is collected.
>
> Actions: 0 = red, 1 = green, 2 = blue. There are 84 trials.

Current observation.

```
[Trial 1 of 84]

Word: RED  (printed in: green)

What is the printed colour? (0=red, 1=green, 2=blue)
Reply with exactly one final line: 'Action: N'.
```

Response-format contract (excerpt). The runtime appends an overriding instruction: "Return only one JSON object with exactly one key, action, whose value is an available integer action. Do not explain the choice or repeat any observation or past record." This JSON contract supersedes the legacy response-style sentence visible in the observation above. Here the permitted values are $\{0, 1, 2\}$; the response contains no additional fields or explanation.

Valid response format (illustrative, not a quoted LLM output).

```
{"action": 1}
```

The harness parses and validates the numeric index before executing it.

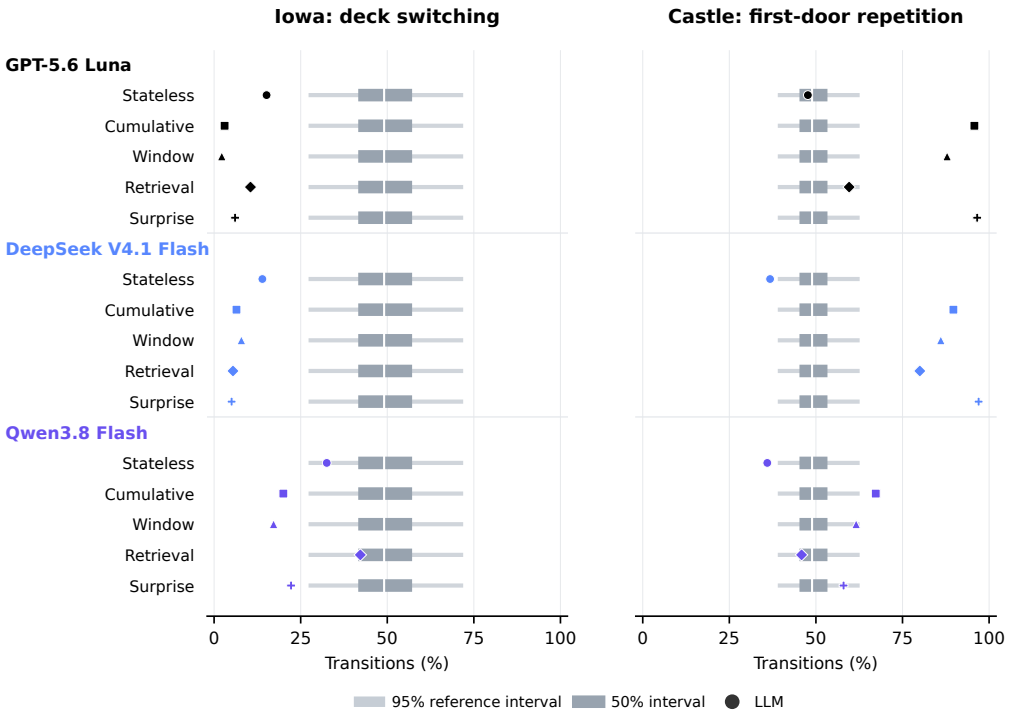

Figure 7: Selected autonomous behavioral checks. Each marker summarizes five LLM episodes; gray bands show matched RNN predictive intervals, with a white tick at the median. Iowa measures switching between decks on consecutive trials. Castle measures repetition of the first-door choice across consecutive training rounds. Colors identify backbones; rows identify memory configurations.

Table 4: Task encodings and frozen training configurations. Configuration entries give encoder/GRU width, followed by learning rate/batch size/maximum epochs/stopping patience. R denotes a continuous head; typed heads depend on response stage.

Task	Encoding; head	Configuration
BART	Pumps, outcomes; 2	16/24;.003/24/20/6
Columbia card	Risk, gain/loss, cards; 2	16/24;.003/24/160/10
Experiential risk	Stimuli, exposures, outcomes; 2	24/32;.003/64/100/10
Ring navigation	Station, goal, visible graph; 2	32/48;.002/64/120/8
Information sampling	Revealed boxes, payoff; 27	20/32;.003/20/80/8
Information demand	Offers, stage, public reveal; typed 2	24/40;.003/64/160/10
Iowa gambling	Deck-specific choice/outcome history; 4	16/24;.003/24/40/6
Weather	Six cues, sun side, phase; 2	24/32;.002/16/18/5
Castle two-step	Stage, prices, past resources; 3	24/32;.003/64/220/12
N-back	Stimulus identity, n, block; 2	16/24;.003/24/40/6
Observe-or-bet	Revealed evidence, block; 3	16/24;.003/24/80/6
Patent race	Endowments, investment, history; 5	20/32;.003/128/80/8
Range learning	Cue identities, phase, feedback; 2	20/28;.003/64/30/5
Reward bandit	Options, prior outcomes; 2	16/24;.003/256/80/10
Risky lottery	Payoffs, probabilities, feedback; 2	24/32;.002/128/80/4
Ultimatum	Offer, trial, offer history; 2	16/24;.003/24/40/6
Retrieve-compare	Cues, phase, revealed outcomes; 2/21	48/48;.002/32/160/15
Reversal learning	Option-specific outcome history; choice	16/24;.003/24/40/6
Momentary mood	Rewards, score, previous mood; R	24/32;.003/128/200/12
Risk/happiness	Offers, event type, past responses; 2+R	32/48;.002/512/12/3
Costly search	Probabilities, cost, revealed payoffs; typed	32/48;.002/64/20/5
Social bandit	Options, visible social cues; 2	24/32;.003/512/60/8
Self–other	Offers, event type, prior feedback; 2/3+R	32/48;.002/64/160/10
Sparse estimation	Event type, revealed amounts; R	32/48;.002/64/100/12
Stroop	Ink/word/candidate relations; choice	16/24;.003/24/40/6

Table 5: Held-out human-response prediction. N: train/validation/test identity groups; T: held-out targets. Accuracy/NLL pairs are unitless/nats. Baselines are the source-designated logistic or grouped linear-choice model (L), or histogram gradient boosting (H). Dashes indicate unavailable comparisons. Continuous errors retain native units.

Task	N	T	GRU acc./NLL	Baseline acc./NLL
BART	1125/141/141	157,919	0.9822/0.0673	L: 0.9819/0.0832
Columbia card	1125/141/140	63,335	0.9321/0.1922	–
Experiential risk	451/56/56	6,160	0.7310/0.5114	H: 0.7093/0.5351
Ring navigation	507/64/63	12,666	0.9392/0.1581	H: 0.9386/0.1692
Information sampling	418/52/52	13,924	0.5753/1.4460	–
Information demand	476/60/59	44,604	0.8776/0.2752	H: 0.8517/0.3262
Iowa gambling	416/52/52	5,685	0.6065/0.9541	L: 0.5147/1.1189
Weather	459/58/57	37,032	0.6200/0.6366	H: 0.5511/0.6870
Castle two-step	493/62/62	12,524	0.6886/0.6623	H: 0.5088/0.9402
N-back	376/47/47	17,990	0.7843/0.4613	L: 0.6504/0.6429
Observe-or-bet	667/84/83	10,375	0.7502/0.5932	–
Patent race	595/74/74	5,920	0.7113/0.7536	H: 0.6767/0.8199
Range learning	640/80/80	19,194	0.5995/0.6140	H: 0.7071/0.5280
Reward bandit	2666/333/333	33,300	0.7499/0.4923	L: 0.7058/0.5569
Risky lottery	4625/578/578	57,800	0.7658/0.4620	–
Ultimatum	535/67/67	1,340	0.9075/0.2194	L: 0.7746/0.4348
Retrieve-compare	539/67/67	28,015	0.8118/0.4349	–
Reversal learning	730/91/91	14,560	0.7887/0.5030	L: 0.7044/0.6527
Momentary mood	963/121/120	7,313	–/–	–
Risk/happiness	37578/4697/4697	259,830	0.7646/0.4876	H: 0.7307/0.5348
Costly search	1486/186/186	52,430	0.7605/0.6679	H: 0.7910/0.5893
Social bandit	1191/149/149	6,705	0.7363/0.5172	H: 0.7289/0.5271
Self–other	502/63/62	3,410	0.8393/0.3736	H: 0.6850/0.5779
Sparse estimation	725/91/90	180	–/–	–
Stroop	91/11/11	924	0.9805/0.1042	L: 0.9805/0.2548

Continuous endpoint	Units	N_{test}	T	GRU MAE	HGB MAE	GRU RMSE	HGB RMSE
Momentary mood	$[-1, 1]$	120	7,313	.1340	.1405	.2211	.2297
Happiness	$[0, 100]$	4,697	103,932	9.6777	11.8931	15.1726	17.1013
Harmful intent	$[0, 100]$	62	62	20.3984	20.8407	28.0039	26.4278
Self-interest	$[0, 100]$	62	62	25.6910	27.2246	33.8311	33.9800
Sparse estimate: post-training	Original slider points	90	90	14.7322	17.0084	23.4832	25.0049
Sparse estimate: post-test	Original slider points	90	90	20.5072	19.5686	25.2041	24.0355

Costly search metrics average typed heads (52,430 targets over 26,215 action events); trajectory scoring uses joint likelihood. Sparse-event slider ranges are 300/500; normalized RMSE is .050912/.052674 for GRU and .054583/.049906 for HGB (post-training/post-test). Retrieve-compare's categorical probability reports also have MAE 14.8996 percentage points over 1,245 reports. Continuous ratings do not enter categorical NLL averages.

Table 6: Mean prompt tokens per decision across the 23 categorical tasks.

Backbone	Stateless	Cumulative	Window	Retrieval	Surprise
GPT-5.6 Luna	2,124	35,026	4,075	4,078	12,470
DeepSeek V4.1 Flash	446	34,223	2,551	2,552	11,146
Qwen3.8 Flash	405	36,269	2,577	2,565	12,653

Table 7: Task-level any-memory completeness contrasts across 23 tasks.

Backbone	Any-memory Δc	Raw p	Holm p
GPT	0.111	0.063419	0.063419
DeepSeek	0.286	0.006033	0.012065
Qwen	0.258	0.000073	0.000512

Table 8: Mean log score S (nats per decision), averaged over five matched seeds. Replay proposals are evaluated on forced histories; closed-loop actions on their own histories.

		Replay S				Closed-loop S			
Task	Backbone	Stateless	Cumul.	Window	Retr.	Stateless	Cumul.	Window	Retr.
Iowa	DeepSeek	-0.738	-0.741	-0.736	-0.756	-0.501	-0.252	-0.305	-0.247
Iowa	Qwen	-1.209	-1.081	-0.809	-0.975	-0.945	-0.581	-0.586	-0.886
Weather	DeepSeek	-0.714	-0.696	-0.697	-0.694	-0.478	-0.403	-0.547	-0.527
Weather	Qwen	-0.704	-0.697	-0.700	-0.707	-0.598	-0.495	-0.624	-0.604
Castle	DeepSeek	-1.902	-0.927	-1.058	-0.930	-1.059	-0.396	-0.570	-0.725
Castle	Qwen	-1.900	-0.977	-1.191	-1.021	-1.140	-0.651	-0.947	-1.004
Social bandit	DeepSeek	-1.015	-0.446	-0.423	-0.544	-0.819	-0.322	-0.319	-0.393
Social bandit	Qwen	-0.974	-0.425	-0.434	-0.507	-0.824	-0.364	-0.403	-0.535

Table 9: Agent–human-policy MAD on the two continuous-only tasks (lower is better). Values are means across completed episodes; parentheses give episode counts. Dashes indicate no completed episodes. Mood and estimation retain their respective native units.

Task	Backbone	Stateless	Cumulative	Window	Retrieval	Surprise
Momentary mood	Luna	–	–	–	–	–
	DeepSeek	0.190 (1)	0.073 (3)	0.070 (3)	–	0.093 (4)
	Qwen	0.251 (1)	0.184 (4)	0.185 (3)	0.248 (2)	0.196 (4)
Sparse estimation	Luna	–	4.3 (5)	11.5 (5)	11.4 (5)	7.5 (5)
	DeepSeek	50.0 (5)	6.3 (5)	67.4 (5)	21.9 (5)	5.8 (5)
	Qwen	69.2 (4)	7.5 (5)	11.5 (5)	19.1 (5)	6.5 (5)