Reasoning aligns language models to human inference

Gonçalo Guiomar[*1 2] **Elia Torre**[*3 2]
Pehuen Moure[2] **Victoria Shavina**[2] **Mario Giulianelli**[4] **Shih-Chii Liu**[2] **Valerio Mante**[3 2]

Abstract

Do language models make decisions under uncertainty like humans do? And if so, what role does *extended reasoning* play in the underlying decision process? We answer this question by introducing an *active probabilistic reasoning* task that cleanly separates *sampling* (actively acquiring evidence) from *inference* (integrating evidence towards a decision). Benchmarking humans and a broad set of contemporary LLMs against optimal reference policies reveals a consistent pattern: extended reasoning is the key determinant of strong performance, driving large gains in inference, while yielding only modest improvements in active sampling. To explain these differences, we fit a behavioral model that captures systematic deviations from optimal Bayesian behavior through interpretable parameter families, placing humans and models in a shared low-dimensional *cognitive space*. The resulting fits show how *reasoning* shifts models toward human-like regimes of evidence accumulation and belief-to-choice mapping, and yield testable predictions about the latent dynamics that might drive each decision. Probing residual-stream activations of an open-weight reasoning model, we find that the geometry of internal representations tracks these predicted dynamics, linking behavior to representational correlates of the fitted latent dynamics.

1. Introduction

Post-training LLMs to produce long intermediate *reasoning traces* before committing to an answer can substantially improve performance on challenging benchmarks, often reaching near-human accuracy (Nye et al., 2021; Reynolds & McDonell, 2021; Wei et al., 2022; Kojima et al., 2022; Wang et al., 2022b; Chen et al., 2025a; Jaech et al., 2024; Guo et al., 2025a; El-Kishky et al., 2025). However, it remains unclear *why* extended reasoning helps and *how* these traces relate to the computations that actually produce the answer. Indeed, the traces can be unreliable and can diverge from the model's internal processes, implying that they do not necessarily describe the true algorithm a model applied in solving a task (Turpin et al., 2023; Lanham et al., 2023; Barez et al., 2025; Chen et al., 2025b; Arcuschin et al., 2025).

More generally, algorithmic-level descriptions of the representations underlying LLM behavior are arguably missing for most current benchmarks (Moskvichev et al., 2023). Yet, a long line of work in cognitive neuroscience suggests that such descriptions could be critically important for any attempts to link a behavior to specific parameters and activations in the underlying circuits (Marr, 2010; Lengyel, 2024). Making explicit the algorithms and representations that lead to a particular behavior can yield *interpretable latent variables* that are often more directly tied to activations than the observable behavior, and thus help bridge the gap between behavior and high-dimensional activations (Lengyel, 2024; Ku et al., 2025).

Such algorithmic descriptions are also needed to evaluate human–LLM alignment and how extended reasoning shapes it. Fine-tuning on human feedback and behavioral datasets can make LLMs reproduce human-like choices and signatures of classic cognitive models (Lake et al., 2017; Peterson et al., 2021; Wei et al., 2021; Bai et al., 2022; Ouyang et al., 2022; Binz & Schulz, 2023a; Momentè et al., 2025; Binz et al., 2025). Yet, performance and behavioral similarity alone may not be sufficient to conclude that LLMs rely on the same latent computations and decision strategies as humans.

In this work, we take inspiration from cognitive neuroscience by developing a behavioral task that isolates two core elements of decision-making under uncertainty: *sampling* and *inference*. Despite its simple structure, our task is demanding, making it a sensitive benchmark for both

[*]Equal contribution [1]ETH AI Center, Zurich, Switzerland [2]Institute for Neuroinformatics, University of Zurich & ETH Zurich, Switzerland [3]ETH Zurich, Switzerland [4]University College London, London, United Kingdom. Correspondence to: Gonçalo Guiomar <goncalo.guiomar@ai.ethz.ch>, Elia Torre <elia.torre@uzh.ch>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

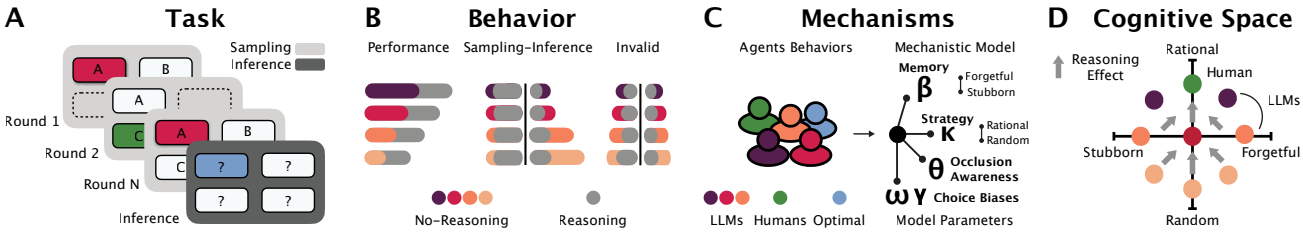

Figure 1. **From task performance to latent cognitive variables. A: Task.** An *active probabilistic reasoning* task in which agents sequentially sample from up to four buttons (A–D), each revealing a binary outcome (RED/GREEN). One button is biased toward RED, while the others are unbiased. During T *sampling* rounds, agents sample evidence among the buttons available. During *inference* rounds, agents indicate which button they believe is biased. An equivalent text-based version is used for LLMs (App. **B.1**). **B: Behavior.** Comparison of human and LLM behavior by *overall performance*, *sampling/inference quality*, and *invalid choices*. These metrics reveal a broad spectrum of performance, with *extended reasoning* improving overall success via enhanced *inference*, while gains in *sampling* remain limited. **C: Mechanisms.** Behavioral model that captures deviations from optimal Bayesian inference using interpretable parameters: **Memory** (β), **Strategy** (m, κ_f), **Choice Bias** (ω, γ), and **Occlusion Awareness** (θ_s, θ_f). **D: Cognitive space.** These parameters define a shared low-dimensional space for humans and LLMs. *Reasoning* shifts LLMs toward human-like *inference* strategies, and tightens the gap in *sampling* strategies.

humans and frontier LLMs. Behavior in our task can be explained in terms of the dynamics of a few, interpretable latent variables, which capture the agents' strategies and how they are shaped by *reasoning*. Concretely, we make the following contributions:

- We introduce an *active probabilistic reasoning task* that disentangles *sampling* from *inference* and is well-suited for process-level analysis of behavior (Section **3**, Fig. **1A**).
- We benchmark humans and a broad set of LLMs against near-optimal reference policies, and find that top *reasoning* models match or exceed humans in *inference quality* but consistently under-perform in *sampling quality* (Sections **4–6**, Fig. **1B**; Fig. **2**).
- We develop a behavioral model that places humans and LLMs in a shared *cognitive space* of interpretable parameters (**Memory**, **Strategy**, **Choice Bias**, **Occlusion Awareness**), showing how *extended reasoning* shifts LLMs toward human-like strategies and produces latent dynamics that yield testable predictions about internal computation (Sections **7**,**8**; Fig. **1C–D**; Fig. **3**).
- We relate behavior to residual-stream activations in *gpt oss 20b*, which encode round-by-round trajectories tracking the latent dynamics predicted by the fitted model (Section **9**; Fig. **4**).

2. Related work

LLMs are increasingly studied as generalist decision-making agents that act in dynamic environments and adapt via *in-context learning* (ICL) and *in-context reinforcement learning* (ICRL) (Brown et al., 2020; Yao et al., 2022; Laskin et al., 2022; Shinn et al., 2023; Liu et al., 2023b;a; Dong et al., 2024; Moeini et al., 2025). A complementary line treats LLMs as candidate models of human cognition, where behavior can potentially be linked to internal representations (Schrimpf et al., 2021; Frank, 2023; Binz & Schulz, 2023b; Aher et al., 2023; Zhao et al., 2025; Loo et al., 2026). These lines intersect in sequential decision-making, e.g. multi-armed bandits (MABs) (Slivkins, 2019; Lattimore & Szepesvári, 2020), where LLMs can leverage feedback to improve choices in-context (Lee et al., 2023; Coda-Forno et al., 2023; Krishnamurthy et al., 2024; Monea et al., 2025; Schubert et al., 2024). Moreover, alignment and fine-tuning on human feedback and behavioral datasets can substantially increase human-likeness, in some cases matching the decision dynamics of traditional cognitive models (Bai et al., 2022; Ouyang et al., 2022; Binz et al., 2024; 2025; Momentè et al., 2025; Su et al., 2025). However, most evaluations emphasize exploration–exploitation or overall performance, offering limited leverage to distinguish strategies that produce similar behavior (Park et al., 2024; Rahn et al., 2024; Nie et al., 2024; Felicioni et al., 2024). Our task departs from classical stochastic MABs by design: instead of conflating information gathering with reward maximization, it separates sampling from inference, which we find to be key to precisely assess alignment between LLMs and humans.

3. Active probabilistic reasoning task

We introduce an *active probabilistic reasoning* task (Fig. **1A**) inspired by classical decision-making and K-armed bandit paradigms (Lai & Robbins, 1985; Daw et al., 2006; Najemnik & Geisler, 2005; Slivkins, 2019; Lattimore & Szepesvári, 2020), but explicitly designed to disentangle *sampling* from *inference*. The task is well-suited for modeling at the algorithmic level, and its structure is "simple" in comparison to many current benchmarks (Jimenez et al., 2023; Sprague et al., 2023; Liu et al., 2023a; Zhou et al., 2023; Mialon et al., 2023; Phan et al., 2025). This simplicity, however, is deceiving, as we show below that *extended rea-*

soning is required for LLMs to achieve high performance in this task, and that even leading models still under-perform compared to humans in key aspects of the task.

Each trial is an independent game with a variable number of *sampling* rounds $T \in \{2, \ldots, 15\}$ followed by a single *inference* round in which the agent must commit to a final decision. On each sampling round, the agent selects one of four buttons (A–D) and observes a binary outcome (RED/-GREEN); at the start of game g, exactly one button is biased towards RED, emitting RED with probability $\alpha_B = 0.9$, while the remaining buttons are unbiased with $\alpha_U = 0.5$. The agent's objective is to identify the biased button from the sampled evidence and report it in the inference round. Formally, the biased button is a latent variable $z_g \in \{1, \ldots, 4\}$; at round t the agent chooses $a_t \in \{1, \ldots, 4\}$ and observes $x_t \in \{0, 1\}$ (with $x = 1$ for RED). To promote active sampling, between 0 and 3 buttons are temporarily occluded on each round.

Although this task resembles a K-armed bandit, it differs in computational goal and information structure: unlike standard bandits where choices yield rewards and agents maximize cumulative return via exploration-exploitation, here performance is determined solely by a one-shot final inference choice. Sampling choices carry no reward but only provide evidence about the single latent hypothesis (*which button is biased*). This separation enables independent assessment of *sampling quality* and *inference quality*.

4. Data collection

We collect behavioral data from both human participants and language models. Human participants played an interactive, graphical version of the task[1]. For language models, we designed an equivalent *text-based* version of the task, in which LLMs interact via token-defined choices under the same instructions provided to the human players (see App. **B.1**).

Humans. We recruited 50 human participants for a 1-hour live in-person competition. All participants provided written informed consent prior to the task and received instructions matching those used to prompt the LLMs (App. **B.1**). Participants played multiple independent games with trial lengths sampled uniformly between 2 and 15 rounds. Of the 50 participants, 46 completed the full protocol of 100 games each, yielding $4,600$ games in total (see App. **F** for individual participant performance profiles).

Large language models. We evaluate a broad set of LLMs spanning state-of-the-art closed-source and open-weight models, covering *dense* and *Mixture-of-Experts* architectures across pre-trained, instruction-tuned and reasoning paradigms (Vaswani et al., 2017; Schulman et al., 2017; Shoeybi et al., 2019; Wei et al., 2021; Ouyang et al., 2022; Wei et al., 2022; Shu et al., 2023; Shao et al., 2024; Cai et al., 2025). Model details are in Figures 2, 3 and App. B.2. We group models into three regimes: *Reasoning* models with a controllable effort parameter, evaluated at *low* and *high* effort (*extended reasoning*, gray bars; Fig. 2A); *CoT* models emitting a *chain-of-thought* trace (max 500 tokens); and *Base* models constrained to a single-token response (see App. B.5, B.6 for breakdowns). Each LLM and condition is evaluated on at least 1,400 games, uniformly spanning 2–15 rounds, for a total exceeding 55,000 games.

5. Optimal agent

To evaluate and compare the performance of LLMs and humans, we introduce an empirical optimal agent that provides an upper bound on task performance in the *finite data setting* of our task. Establishing the optimal strategy for *inference* is straightforward: an agent should select the button with the highest posterior probability of being biased, following a *Maximum-A-Posteriori* (MAP) decision rule (Griffiths et al., 2008). This *optimal inference* strategy is defined as follows; button outcomes are generated by Bernoulli emissions, with the emission probabilities $\{\alpha_B, \alpha_U\}$ for the biased and unbiased buttons respectively. A single game g of T rounds consists of a sequence of choice-evidence pairs $(a_{1:T}, x_{1:T})$ together with a latent context variable $k \in \{1, \ldots, K\}$. We represent beliefs by the posterior vector $(\boldsymbol{p}_t)_k \in \Delta^K$. We also define the likelihoods over the latent variable k and emission rates $\{\alpha_B, \alpha_U\}$ by $(\boldsymbol{L}_t)_k \in \Delta^K$.

By assuming a uniform prior, we can write down the recursive normative posterior via Bayes' rule as (see App. K):

$$\boldsymbol{p}_t = \frac{\boldsymbol{p}_{t-1} \odot \boldsymbol{L}_t}{\mathbf{1}^\top (\boldsymbol{p}_{t-1} \odot \boldsymbol{L}_t)}, \qquad (\boldsymbol{p}_0)_k = \tfrac{1}{4}. \tag{1}$$

where $\odot$ denotes the Hadamard product. The MAP *inference* policy at round t selects the hypothesis with maximum posterior and is given by:

$$a_t^{\mathrm{MAP}} := \arg\max_{k \in \{1,\ldots,4\}} (\boldsymbol{p}_t)_k. \tag{2}$$

Determining the optimal *sampling* strategy is more involved because the agent chooses which evidence to acquire under random occlusions. For empirical benchmarking (Fig. **2**) and loss decomposition, we use a high-performing RL reference policy evaluated under the same regime as humans and LLMs. The RL reference uses a PPO-trained sampling policy (Schulman et al., 2017) together with the same Bayesian MAP rule for the final decision (App. **C**). Nonetheless, this reference policy matches the theoretical Dynamic Programming optimum (App. **B.9**).

[1] https://ai.trt-bench.org

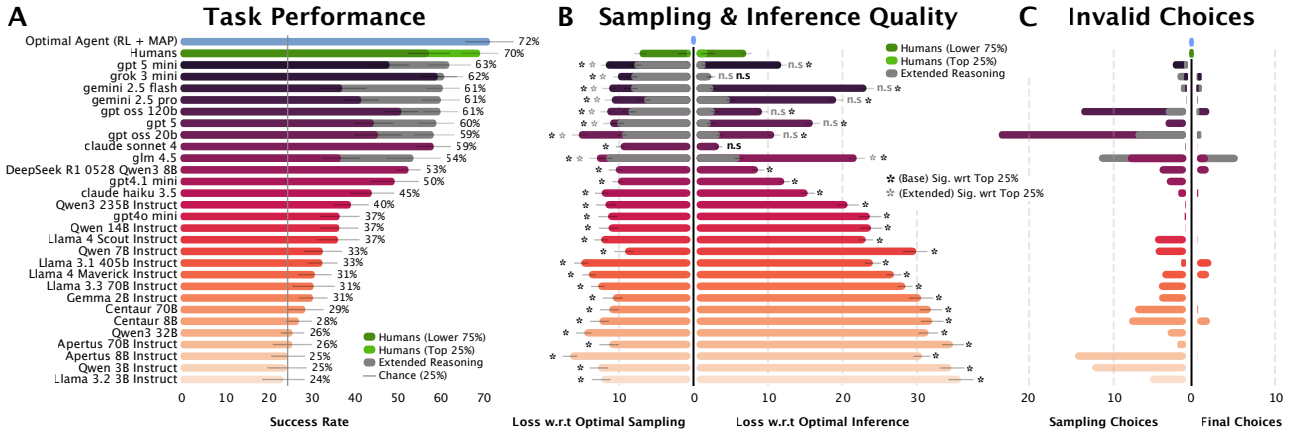

Figure 2. **Comparing human and LLM behavior. A: Task performance.** Mean success rate across trial lengths $T \in \{2, \ldots, 15\}$. Shown are humans (green; bottom 75% and top 25%), the optimal agent (PPO+MAP; light blue), and, where available, *Extended Reasoning* (gray overlays). Error bars denote standard deviations across trial-cluster means under a uniform distribution over round count. The vertical line marks chance (25%). **B: Sampling and inference quality.** Sampling loss (left) and inference loss (right), both relative to the optimal agent (lower is better). Inference loss is the gap to a counterfactual agent that keeps the same sampled evidence but applies MAP at the final decision; sampling loss is the gap between this counterfactual MAP agent and the reference agent, isolating suboptimal evidence acquisition. Error bars show bootstrap standard deviations ($B = 1{,}000$). Asterisks indicate permutation tests against the human top-25% baseline ($*$: $p < 0.0001$; n.s.: not significant; Bonferroni-corrected); black denotes base, gray extended reasoning. Across the 8 models tested in both conditions, reasoning reduced inference loss by 10.9 pp on average versus 3.0 pp for sampling loss (95% bootstrap CI for the difference: $[+3.5, +12.4]$ pp; paired permutation $p = 0.015$); see App. **B.3**, **B.4** for per-model statistics. **C: Invalid choices.** Fraction of invalid choices during sampling (left) and at the final inference decision (right). Invalid choices include selecting occluded options, producing tokens outside A–D, or failing to respond; humans cannot make invalid choices in the graphical interface. Invalid choices are more frequent during sampling and decrease with reasoning.

6. Comparing human & LLM behavior

To compare human and LLM performance, we employ four complementary metrics that capture distinct aspects of behavior: (1) overall task performance, (2) sampling quality, (3) inference quality, and (4) invalid choice rates.

Overall task performance. In Figure **2A**, we quantify *overall task performance* as the average *success rate* across all trial lengths ($T \in \{2, \ldots, 15\}$), i.e., the fraction of games in which the agent correctly identifies the biased button at the inference round. Our task induces a broad distribution of performance across both humans and models with clear gains from *extended reasoning* (gray vs. colored bars). On average, human participants achieve $\simeq 61\%$ success, comparable to the best-performing LLM (*gpt 5 mini*); however, the top quartile of humans exceeds this level by $\sim 7\%$, highlighting wide human variability and a remaining gap between top models and skilled human players. Across models, *reasoning* LLMs outperform *non-reasoning* counterparts with success rates rising as games lengthen, while *non-reasoning* models stay flat, failing to leverage the additional evidence (App. **E**).

Sampling & inference quality. In Figure **2B**, we quantify *inference quality* by computing the performance gap between each agent and a counterfactual agent that retains the same sampling strategy but performs MAP inference. This difference quantifies the performance loss attributable solely to suboptimal *inference* (**2B** right, lower is better). We also evaluate *sampling quality*, by comparing this counterfactual agent (which combines the agent's sampling strategy with MAP inference) against the optimal RL agent defined above. Since both agents share the optimal MAP inference rule, any remaining performance gap reflects the loss due to suboptimal *sampling* strategy (Fig. **2B** left, lower is better). We find that *inference quality* varies broadly across LLMs and is strongly correlated (negatively, because the metric is a loss) with overall performance ($r = -0.99$), whereas *sampling quality* is more homogeneous across LLMs and a weaker predictor of performance ($r = -0.47$; $r = -0.70$ if we include *extended reasoning* variants). Performance gains from *extended reasoning* are driven mainly by improved inference: across the 8 models tested under both reasoning conditions, reasoning reduced inference loss by 10.9 pp on average versus only 3.0 pp for sampling loss (95% bootstrap CI: $[+3.5, +12.4]$ pp; paired permutation $p = 0.015$). Frontier reasoning models match or exceed the top-25% humans in *inference quality* (see App. **B.3**, **B.4**); in contrast, *sampling quality* remains a weakness: the best reasoning models only approach the lower 75% of human participants, while the top 25% still substantially outperform all LLMs (left-hand side of Fig. **2B**). Notably, this dissociation is unlikely to reflect a pure compute-budget effect: under *high-reasoning*, sampling rounds consume far

more tokens than inference decisions (∼8k–12k vs. ∼2k–5k for *gpt oss 20b*; App. **B.7**), yet sampling benefits the least. Representative traces suggest that inference gains instead reflect explicit evidence-integration steps (App. **B.8**).

Invalid choice rates. Finally, in Figure **2C**, we quantify *invalid choices* as the proportion of choices in which agents produce invalid responses during either *sampling* or *inference*. Invalid choices include selecting occluded buttons, generating output tokens outside the valid button vocabulary (A–D), or failing to produce a response altogether. By design, human participants cannot produce invalid choices in the graphical interface, and therefore invalid-choice rates are computed only for LLMs. Overall, we do not observe a statistically significant correlation between invalid choice rates and LLM performance; however, invalid choices are consistently reduced under *extended reasoning*. Invalid responses tend to occur more often during *sampling* than during *inference*, potentially affected by choice occlusions during *sampling rounds*.

7. A behavioral model of latent cognitive variables

The behavioral metrics above reveal systematic differences across LLMs, but not the processes that generate them. We therefore fit a compact behavioral model to test how reasoning changes *sampling* and *inference* strategies. The model is motivated by Bayesian ideal-observer and probabilistic cognition frameworks (Knill & Pouget, 2004; Sanborn et al., 2010), and by views of *in-context learning* as implicit inference over latent context variables (Xie et al., 2022; Panwar et al., 2024; Zhang et al., 2023). On this normative backbone, it allows systematic deviations in **memory**, **strategy**, **choice bias/inertia**, and **occlusion awareness**. These explanations remain behavioral, but provide a structured prior over possible mechanisms. We use the fitted parameters to place agents in a shared behavioral parameter space (Figs. **3A–G**), which guides the representational analyses in Sec. **9**.

The fitted parameters explained below give a compact and validated description of the processes agents may implement in the task (Fig. **3**; full values in App. **K.9**). Fitted-model simulations reproduce the main human and LLM behavioral metrics (App. **G**), and held-out choices are predicted better than by an agent-specific frequency baseline for every agent-condition pair (27/27; mean ΔeLPD/game $= 6.67$, median $= 6.99$, all adjusted $p < 0.001$; median improved-game fraction $= 98.5\%$; App. **K.6**). The parameters are also interpretable and stable: each produces distinct behavioral signatures (App. **K.8**), recovers reliably from simulated data and across fits (App. **K.7**), lies in well-behaved likelihood landscapes (App. **K.7**), and shows negligible repeated-fold train–test gaps across 39 model-condition fits (mean test-train NLL gap $= 0.0015$).

Memory (β). As in sequential sampling models (Busemeyer & Townsend, 1993; Usher & McClelland, 2001), choices depend on a non-optimal *memory* $\boldsymbol{h}_t \in \mathbb{R}^K$ (Gold & Shadlen, 2007; Hogarth & Einhorn, 1992; Ma et al., 2023). Let $\boldsymbol{p}_t$ be the Bayesian posterior after round t, set $\boldsymbol{h}_0 = \mathbf{0}$, and define centered log-posterior increments

$$\Delta \boldsymbol{h}_t := C(\log \boldsymbol{p}_t) - C(\log \boldsymbol{p}_{t-1}), \tag{3}$$

with $C(\boldsymbol{x}) = \boldsymbol{x} - \langle \boldsymbol{x} \rangle \mathbf{1}$, and with which we can define our memory integration process by applying a temporal kernel to previous increments:

$$\boldsymbol{h}_t = \sum_{u=1}^{t} w_{t,u}(\beta) \Delta \boldsymbol{h}_u, \tag{4}$$

Here, $w_{t,u}(\beta) = \exp[-\beta(t-u)]$ for $\beta > 0$, $w_{t,u}(\beta) = 1$ for $\beta = 0$, and $w_{t,u}(\beta) = \exp[\beta(u-1)]$ for $\beta < 0$. Thus $\beta = 0$ recovers perfect accumulation, $\beta > 0$ produces *forgetful* recency weighting, and $\beta < 0$ produces *stubborn* primacy weighting where earlier evidence dominates.

Strategy (κ_s, κ_f). We map beliefs to actions with a softmax/logit choice rule, following probabilistic-choice and bounded-rational decision models (Luce, 1959; McFadden, 1974; Ortega & Braun, 2013) using time-varying *sampling* gain $\kappa_s(t)$ and a static *inference* gain parameter. For *sampling*, as the outcome at round t has not yet been observed, sampling depends on the pre-choice memory $\boldsymbol{h}_{t-1}$. Let $o_u \in \{0, 1\}$ indicate whether round u yielded a valid observation, and define

$$n_t := 1 + \sum_{u<t} o_u, \qquad \kappa_s(t) := m \, \frac{n_t}{T}, \tag{5}$$

where T is the number of sampling rounds in the game. The gain m controls how quickly sampling shifts from exploratory to evidence-guided as usable evidence accumulates (Daw, 2011; Gershman, 2019); κ_f controls final-choice sharpness, approaching the MAP rule as $\kappa_f \to \infty$.

Choice Bias ($\boldsymbol{\omega}, \gamma$). To dissociate evidence-driven choices from systematic preferences, we include choice-bias vectors $\{\boldsymbol{\omega}_s, \boldsymbol{\omega}_f\} \in \Delta^K$ for sampling and final choices, together with sampling action inertia γ (Gershman, 2020; Wei et al., 2024; Jiang et al., 2024). Inertia adds $\gamma \boldsymbol{i}_t$ during sampling, where $\boldsymbol{i}_t \in \{0, 1\}^K$ is one-hot for the most recent valid sampling choice and zero if none exists.

Occlusion Awareness (θ_s, θ_f). During sampling, the available action set changes from round to round. We model sensitivity to this constraint by adding the log-score penalty $-\theta_s(\mathbf{1} - \boldsymbol{v}_t)$, where $\boldsymbol{v}_t \in \{0, 1\}^K$ indicates which cues

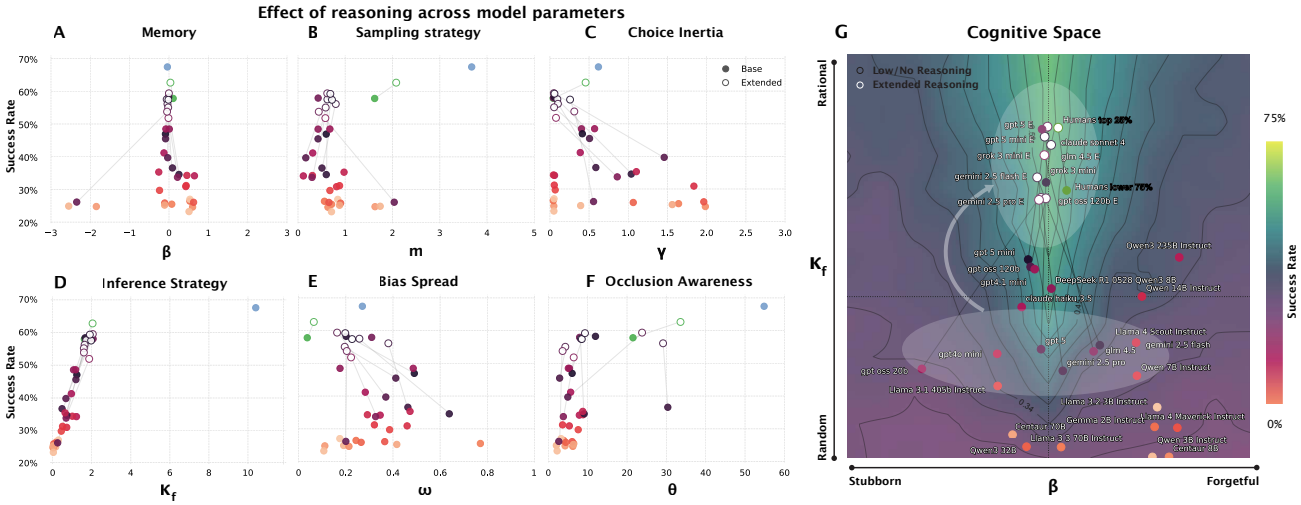

Figure 3. **Reasoning aligns language models to human inference.** Each point shows the 5-fold CV fit average of one fitted agent (see fit quality details in App. **K.5**, with LLMs colored by their position in the benchmark in Fig. **2**, humans in green, and the optimal agent in blue; filled markers denote *base* and open markers denote *extended reasoning*, with paired conditions connected by light gray lines. **A**: **Memory** β versus task performance, spanning *stubborn* ($\beta < 0$) to *forgetful* ($\beta > 0$) evidence integration. **B**: **Sampling strategy scale** m versus performance. m controls the linear temporal sampling gain $\kappa_s(t)$. **C**: **Choice inertia** γ versus performance, capturing *perseveration* on the most recent valid sampling choice. **D**: **Inference sharpness** κ_f versus performance, controlling how decisively memory is mapped to the final inference policy. **E**: **Bias spread** summarizes systematic choice preference strength as the mean range of sampling and inference bias vectors. **F**: **Occlusion awareness** θ versus performance, capturing avoidance of invalid or occluded choices. **G**: **Cognitive Space**. *Reasoning* models move in the cognitive subspace defined by β, κ_f towards the human-level inference regime with the heatmap representing the task performance landscape for a grid sweep of the behavioral model with the rest of the parameters fixed at their optimal values.

are available. This is a soft behavioral analogue of invalid-action masking in reinforcement learning (Huang & Ontañón, 2022): as $\theta_s \to \infty$, unavailable choices receive zero probability, while smaller θ_s allows lapses or failures to respect the mask.

Policy. The stochastic policy is obtained by adding memory, choice bias, action inertia, and occlusion-awareness terms yielding a multinomial-logit/softmax policy (Luce, 1959; McFadden, 1974) and is given by:

$$\boldsymbol{\eta}_t^s = \underbrace{\kappa_s(t,T)}_{\text{strategy}}\underbrace{\boldsymbol{h}_{t-1}}_{\text{memory}} + \underbrace{\log \boldsymbol{\omega}_s}_{\text{bias}} + \underbrace{\gamma \boldsymbol{i}_t}_{\text{inertia}} - \underbrace{\theta_s(\mathbf{1} - \boldsymbol{v}_t)}_{\text{occlusion penalty}}, \quad (6)$$

$$\boldsymbol{\eta}^f = \underbrace{\kappa_f}_{\text{strategy}}\underbrace{\boldsymbol{h}_T}_{\text{memory}} + \underbrace{\log \boldsymbol{\omega}_f}_{\text{bias}}, \quad (7)$$

Conditional on producing a cue response, the agent chooses the *sampling* or *inference* action through $\boldsymbol{\pi}_t^s = \sigma(\boldsymbol{\eta}_t^s)$ and $\boldsymbol{\pi}^f = \sigma(\boldsymbol{\eta}^f)$, respectively. Invalid or missing final responses are handled by θ_f in the behavioral likelihood. The loss function used, full model derivations and fitting details are shown in App. **K**. A large set of different models were tested, with the summarized model search detailed in App. **K.10**.

8. Reasoning shortens the inference gap between LLMs and Humans

We next use the fitted parameters diagnostically, asking which components of the behavioral model covary with success across humans, LLMs, and the optimal reference agent. Figure **3** relates each parameter to task performance. The strongest gradients appear in inference-related parameters. High-performing agents have memory parameters (β) close to zero (Fig. **3A**), consistent with near-perfect evidence accumulation, whereas low-performing models often occupy more extreme *stubborn* ($\beta < 0$) or *forgetful* ($\beta > 0$) regimes; β is interpreted cautiously for near-random agents, where it is weakly identifiable (App. **K.7**). Performance also increases sharply with *inference* strategy (κ_f; Fig. **3D**): low-performing models have κ_f near zero, implying weak use of accumulated evidence, whereas the best LLMs approach a sharper, more MAP-like *inference* policy. These trends explain why *extended reasoning* primarily improves *inference quality* in the behavioral decomposition (Fig. **2B**).

Sampling-related parameters show a more mixed pattern. The sampling gain (m) increases for many high-performing agents (Fig. **3B**), but remains below the *optimal agent*, consistent with the persistent *sampling quality* gap (Fig. **2B**, left). Choice inertia (γ) is largest among several low-performing models (Fig. **3C**), indicating *perseveration* on the most recent valid sampling action. Bias spread, the

mean range of the *sampling* and *inference* bias vectors, varies substantially across models (Fig. **3E**); larger values indicate stronger evidence-independent preferences. Occlusion awareness (θ) also varies widely (Fig. **3F**), matching differences in invalid-choice rates across LLMs (Fig. **2C**).

Paired *base* and *extended* conditions show how *reasoning* shifts these latent strategies (Fig. **3**, filled vs. open markers connected by gray lines). *Extended reasoning* usually improves LLM performance while moving models toward the human-like region of parameter space: β moves closer to zero, κ_f increases, γ tends to decrease, and θ often increases. The shift in m is smaller, indicating that *reasoning* improves *active sampling* less reliably than *final inference*. Bias spread is also only partly reduced: some models become less biased, but fitted biases remain heterogeneous even among high-performing agents. Thus, *reasoning* improves performance mainly by sharpening evidence integration and final belief-to-choice mapping (Fig. **3G**), while leaving sampling and bias-related deviations only partly corrected (parameter details in App. **G**).

9. Activation geometry tracks model-predicted latent dynamics

The behavioral model provides hypotheses about the latent dynamics that may guide agents' choices, making concrete predictions about choice evolution across rounds via the posterior $\boldsymbol{p}_t$, memory state $\boldsymbol{h}_t$, and policy $\boldsymbol{\pi}_t$ (Eqs. **1–6**). We do not try to identify the fitted variables themselves inside the model; instead, we ask whether internal choice representations change in ways predicted by the fitted trajectories (Wang et al., 2022a; Gurnee & Tegmark, 2023; Bereska & Gavves, 2024; Sharkey et al., 2025). Whereas past efforts to interpret *reasoning* have been hampered by the open-ended space of long reasoning traces (Bogdan et al., 2025; Macar et al., 2025), the behavioral model narrows this space to a small set of testable predictions. As a case study, we probe residual-stream activations of *gpt oss 20b* under *low-* and *high-reasoning* and ask whether activation trajectories move between choice-associated regions as the behavioral model predicts.

Distinct latent dynamics under low and high reasoning. Fig. **4A–C** illustrates the distinct latent dynamics for a single game played by *gpt oss 20b* under *high-* (Fig. **4B**) and *low*-reasoning effort (Fig. **4C**). The evidence sequence (Fig. **4A**, top) is identical across conditions, as it is the Bayesian posterior $\boldsymbol{p}_t$ (Fig. **4A**). Around round 11, $\boldsymbol{p}_t$ flips from favoring *button C* to *button A* as evidence accumulates. Under *high reasoning*, the fitted memory state $\boldsymbol{h}_t$ closely tracks this switch ($\beta \approx 0$) and the resulting policy $\boldsymbol{\pi}_t$ becomes sharply peaked on the correct option ($\kappa_f \gg 1$) (Fig. **4B**). Under *low reasoning*, $\boldsymbol{h}_t$ instead remains *stubborn* ($\beta < 0$), continuing to favor *C* despite the new evidence, and $\boldsymbol{\pi}_t$ stays diffuse ($\kappa_f \approx 0$) (Fig. **4C**). The behavioral model therefore predicts a specific neural signature: under *high reasoning*, internal representations should reorganize around *A* as evidence accumulates, whereas under *low reasoning* they should remain anchored near *C*.

Choices are linearly decodable from residual-stream activations. To test this prediction, we collected residual-stream activations from *gpt oss 20b* ($d = 2880$, 24 blocks) while it played 50 games per *reasoning* condition, taking the average of the last 5 *reasoning trace* token activations at each round and each layer (see App. **I**). Following standard probing methodology (Alain & Bengio, 2016; Belinkov, 2022; Gurnee & Tegmark, 2023), we fit ridge regression probes with leave-one-out selection of the regularization strength to predict the agent's choice at each round, with held-out R^2 averaged over 5-fold cross-validation. Choices are decoded with $R^2 > 0.8$ in both *reasoning* conditions, peaking in early-to-mid layers (Fig. **4D**). A *swapped control* (SC) yields R^2 near zero (App. **I**).

Activation trajectories track the model-predicted dynamics. Establishing that choices are linearly readable in early layers, we asked whether the round-by-round *reasoning stream* activation trajectory reorganizes as the fitted latents predict for the example game in Fig. **4A**. We ran PCA on activations from layer 1 (near the peak of probe R^2), averaged across games within each round, choice, and reasoning level (App. **J**); the top four components capture more than 60% of the variance under both *reasoning levels*. For each choice, we computed a *centroid* as the mean activation across rounds in the original 2880-dimensional space, and measured each round's *Euclidean distance* to all four centroids, reported as a *softmax-normalized* similarity vector (Fig. **4B–C**, bottom row). The example game's trajectory was then projected into the PCA basis and colored by the closest centroid at each round (Fig. **4E–F**).

The resulting geometry recapitulates the behavioral model's prediction. Under *high reasoning* (Fig. **4E**), the trajectory begins near the *C* centroid and tracks the posterior flip with a slight lag: the closest centroid first switches away from *C* at $t = 12$, transiently passes through *B*, and stabilizes on *A* from $t = 14$ onward. On the contrary, under *low reasoning* (Fig. **4F**) it remains anchored near *C* throughout the game. This pattern is also visible in the per-round centroid-distance heatmaps (Fig. **4B–C**, bottom row): the *distance vector* closely resembles the fitted latent variables in each condition, providing a correspondence between the behavioral model and a quantity computed from activations. Although our analyses are correlational and cannot establish the causal subspaces underlying LLM decisions, targeted interventions could test the role of these latent variables in

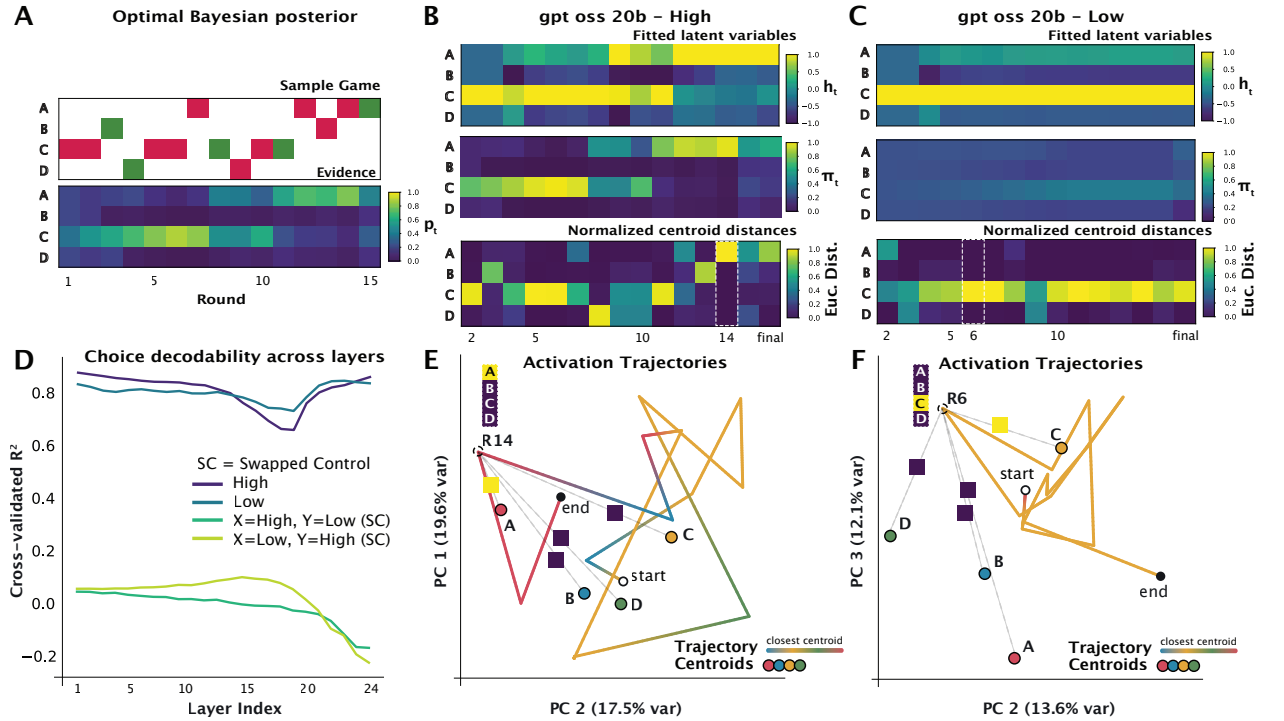

Figure 4. **Activation geometry is consistent with latent dynamics predicted by the behavioral model. A: Optimal Bayesian posterior.** Evidence sequence (top) and Bayesian posterior $\boldsymbol{p}_t$ (bottom) for an example game. **B–C: Latent-variable dynamics and centroid distances** for *gpt oss 20b* under *high-* (**B**) and *low-reasoning* (**C**). For each condition, we show the memory state $\boldsymbol{h}_t$ (top), policy $\boldsymbol{\pi}_t$ (middle), and softmax-normalized centroid-distance vector computed from the model's activations (bottom; see App. **J**). Under *high reasoning*, $\boldsymbol{h}_t$ tracks the posterior switch from *C* to *A* ($\beta \approx 0$) and $\boldsymbol{\pi}_t$ becomes sharply peaked ($\kappa_f \gg 1$); under *low reasoning*, $\boldsymbol{h}_t$ remains stubborn ($\beta < 0$) and $\boldsymbol{\pi}_t$ stays diffuse ($\kappa_f \approx 0$). The activation-derived distance vectors recapitulate these dynamics. Dashed boxes mark the example rounds shown in panels **E–F**. **D: Choice decodability across layers.** Cross-validated R^2 of ridge probes predicting the agent's choice from residual-stream activations (*high* in purple, *low* in teal). *Swapped controls* (SC; green/yellow) train probes on one condition and evaluate on the other; SC R^2 collapses to near zero. **E–F: Activation trajectories** of the example game projected into the layer-1 PCA basis under *high* (**E**) and *low* (**F**) *reasoning*. Filled circles mark the per-choice centroids; the trajectory is colored by the closest centroid at each round. Under *high reasoning*, the trajectory begins near *C* and migrates toward *A*; under *low reasoning*, it remains anchored near *C* throughout the game. Pairwise PC views are shown in App. Fig. **11**.

the model's computations and serve as probes for understanding cognitive representations in biological organisms (Ravichander et al., 2021).

10. Discussion

We asked whether language models form beliefs and make decisions under uncertainty like humans, and what role *extended reasoning* plays. By separating *sampling* from *inference*, fitting a compact *behavioral model*, and probing internal representations we could ask not only *whether* models behave like humans, but also *how* their algorithms differ.

Across LLMs, *reasoning* is the main driver of strong performance, but its effects are selective: it reduces *inference loss* roughly four times more than *sampling loss*, despite *sampling* consuming more tokens, arguing against a simple token-budget account. At the behavioral-model level, *reasoning* sharpens evidence integration and belief-to-choice mapping ($\beta \approx 0$, larger κ_f, reduced *bias* and *perseveration*) while leaving active information acquisition systematically suboptimal. Notably, models fine-tuned on human behavioral data (Binz et al., 2025) perform near chance and occupy the stubborn/low-κ_f region of cognitive space, far from human participants. This suggests that imitating aggregate human behavior on prior datasets does not transfer to the algorithmic regime humans use for active probabilistic inference.

The fitted latent variables also yield testable hypotheses for the underlying computations. In *gpt oss 20b*, residual-stream geometry recapitulates them: under *high reasoning*, representations migrate from an initial favorite toward the correct option as evidence accumulates; under *low reasoning*, they remain anchored in the *stubborn-integration* regime the fits identified. Algorithmic-level descriptions thus serve as scaffolding for implementation-level analysis, narrowing the otherwise open-ended space of computations a long *reasoning trace* might perform.

Together, these results point toward evaluations that measure capability while constraining mechanism, asking not only *whether* models solve a task, but *how*. The persistent *sampling* gap is a particularly informative residual: it marks a class of computations that current reasoning training does not reach, one that *process-level* benchmarks of this kind are well-suited to expose.

11. Reproducibility Statement

To ensure reproducibility of our results, all code used in our experiments, along with detailed instructions for setup and execution, is available at: https://drive.google.com/drive/folders/17tQxO02lLN1VpwbOF_IIiM9oSm8DmRik. Additionally, the active probabilistic reasoning task used for data collection is accessible at https://ai.trt-bench.org.

12. Impact Statement

This paper presents work whose goal is to improve how reasoning in language models is evaluated, compared, and interpreted. A potential positive impact is that tasks which disentangle evidence acquisition from inference may support more transparent benchmarking, better diagnosis of failure modes, and more targeted alignment or safety interventions when models behave in human-like yet systematically imperfect ways.

Potential negative impacts also exist. Behavioral analyses of the kind we study could be used to optimize systems for more persuasive, manipulative, or deceptive decision policies, and human–LLM similarity on a narrow task could be over-interpreted as evidence of broader human alignment. The human-participant component of the work likewise requires careful handling of consent, privacy, and data governance, so only task behavior—rather than personally identifying information—should be shared with released artifacts.

References

Aher, G. V., Arriaga, R. I., and Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pp. 337–371. PMLR, 2023.

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Anthropic. Claude opus 4 & claude sonnet 4 — system card. Technical report, Anthropic, 2025. Accessed 2025-09-23.

Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., and Conmy, A. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Barez, F., Wu, T.-Y., Arcuschin, I., Lan, M., Wang, V., Siegel, N., Collignon, N., Neo, C., Lee, I., Paren, A., et al. Chain-of-thought is not explainability. *Preprint, alphaXiv*, pp. v1, 2025.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Bereska, L. and Gavves, E. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

Binz, M. and Schulz, E. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*, 2023a.

Binz, M. and Schulz, E. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*, 2023b.

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., Griffiths, T. L., Haridi, S., Jagadish, A. K., Ji-An, L., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., Modirshanechi, A., Nath, S. S., Peterson, J. C., Rmus, M., Russek, E. M., Saanum, T., Scharfenberg, N., Schubert, J. A., Buschoff, L. M. S., Singhi, N., Sui, X., Thalmann, M., Theis, F., Truong, V., Udandarao, V., Voudouris, K., Wilson, R., Witte, K., Wu, S., Wulff, D., Xiong, H., and Schulz, E. Centaur: a foundation model of human cognition, 2024. URL https://arxiv.org/abs/2410.20268.

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., Griffiths, T. L., Haridi, S., Jagadish, A. K., Ji-An, L., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., Modirshanechi, A., Nath, S. S., Peterson, J. C., Rmus, M., Russek, E. M., Saanum, T., Scharfenberg, N., Schubert, J. A., Buschoff, L. M. S., Singhi, N., Sui, X., Thalmann, M., Theis, F., Truong, V., Udandarao, V., Voudouris, K., Wilson, R. C., Witte, K., Wu, S., Wulff, D., Xiong, H., and Schulz, E. A foundation model to predict and capture human cognition. *Nature*, 2025. doi: 10.1038/s41586-025-09215-4.

Bogdan, P. C., Macar, U., Nanda, N., and Conmy, A. Thought anchors: Which llm reasoning steps matter? *arXiv preprint arXiv:2506.19143*, 2025.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Busemeyer, J. R. and Townsend, J. T. Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100(3): 432–459, 1993. doi: 10.1037/0033-295X.100.3.432.

Cai, W., Jiang, J., Wang, F., Tang, J., Kim, S., and Huang, J. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

Chen, Q., Qin, L., Liu, J., Peng, D., Guan, J., Wang, P., Hu, M., Zhou, Y., Gao, T., and Che, W. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025a.

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025b.

Coda-Forno, J., Binz, M., Akata, Z., Botvinick, M., Wang, J. X., and Schulz, E. Meta-in-context learning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Comanici, G. et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv*, 2025.

Daw, N. D. Trial-by-trial data analysis using computational models: (tutorial review). In Delgado, M. R., Phelps, E. A., and Robbins, T. W. (eds.), *Decision Making, Affect, and Learning: Attention and Performance XXIII*. Oxford University Press, 2011. doi: 10.1093/acprof:oso/9780199600434.003.0001. URL https://doi.org/10.1093/acprof:oso/9780199600434.003.0001.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., and Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature Neuroscience*, 9(6):867–873, 2006. doi: 10.1038/nn1743.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 1107–1128, 2024.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

El-Kishky, A., Wei, A., Saraiva, A., Minaiev, B., Selsam, D., Dohan, D., Song, F., Lightman, H., Clavera, I., Pachocki, J., et al. Competitive programming with large reasoning models. *arXiv preprint arXiv:2502.06807*, 2025.

Felicioni, N., Maystre, L., Ghiassian, S., and Ciosek, K. On the importance of uncertainty in decision-making with large language models. *Transactions on Machine Learning Research*, 2024.

Frank, M. C. Openly accessible llms can help us to understand human cognition. *Nature Human Behaviour*, 7(11): 1825–1827, 2023.

Gemma Team, Google DeepMind. Gemma 3 technical report. *arXiv*, 2025.

Gershman, S. J. Uncertainty and exploration. *Decision*, 6 (3):277–286, 2019. doi: 10.1037/dec0000101.

Gershman, S. J. Origin of perseveration in the trade-off between reward and complexity. *Cognition*, 204:104394, November 2020. doi: 10.1016/j.cognition.2020.104394.

Gold, J. I. and Shadlen, M. N. The neural basis of decision making. *Annual Review of Neuroscience*, 30:535–574, 2007. doi: 10.1146/annurev.neuro.29.051605.113038.

Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. Bayesian models of cognition. In Sun, R. (ed.), *The Cambridge Handbook of Computational Psychology*, pp. 59–100. Cambridge University Press, 2008. doi: 10.1017/CBO9780511816772.006.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.

Guo, D. et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv*, 2025b.

Gurnee, W. and Tegmark, M. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

Hernández-Cano, A. et al. Apertus: Democratizing open and compliant llms for global language environments. *arXiv*, 2025.

Hogarth, R. M. and Einhorn, H. J. Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24(1):1–55, 1992. doi: 10.1016/0010-0285(92)90002-J.

Huang, S. and Ontañón, S. A closer look at invalid action masking in policy gradient algorithms. In *Proceedings of the Thirty-Fifth International Florida Artificial Intelligence Research Society Conference*, 2022. doi: 10.32473/flairs.v35i.130584.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Jiang, B., Xie, Y., Hao, Z., Wang, X., Mallick, T., Su, W. J., Taylor, C. J., and Roth, D. A peek into token bias: Large language models are not yet genuine reasoners. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4722–4756, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.272.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

Knill, D. C. and Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719, 2004. doi: 10.1016/j.tins.2004.10.007.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Krishnamurthy, A., Harris, K., Foster, D. J., Zhang, C., and Slivkins, A. Can large language models explore in-context? *Advances in Neural Information Processing Systems*, 37:120124–120158, 2024.

Ku, A., Campbell, D., Bai, X., Geng, J., Liu, R., Marjieh, R., McCoy, R. T., Nam, A., Sucholutsky, I., Veselovsky, V., et al. Using the tools of cognitive science to understand large language models at different levels of analysis. *arXiv preprint arXiv:2503.13401*, 2025.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1): 4–22, 1985. doi: 10.1016/0196-8858(85)90002-8.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Laskin, M., Wang, L., Oh, J., Parisotto, E., Spencer, S., Steigerwald, R., Strouse, D., Hansen, S., Filos, A., Gazeau, M., et al. In-context reinforcement learning with algorithm distillation. *arXiv preprint arXiv:2210.14215*, 2022.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lee, J. N., Xie, A., Finn, C., Pacchiano, A., Chandak, Y., Nachum, O., and Brunskill, E. Supervised pretraining can learn in-context reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Lengyel, M. Marr's three levels of analysis are useful as a framework for neuroscience. *The Journal of Physiology*, 602(9):1911–1914, 2024. doi: 10.1113/JP279549. Epub 2024 Apr 16.

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023a.

Liu, Z., Hu, H., Zhang, S., Guo, H., Ke, S., Liu, B., and Wang, Z. Reason for future, act for now: A principled framework for autonomous llm agents with provable sample efficiency. *arXiv preprint arXiv:2309.17382*, 2023b.

Loo, A., Pavlick, E., and Feiman, R. Llms model how humans induce logically structured rules. *Journal of Memory and Language*, 146:104675, 2026.

Luce, R. D. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley & Sons, New York, 1959.

Ma, W. J., Kording, K. P., and Goldreich, D. *Bayesian Models of Perception and Action: An Introduction*. The MIT Press, Cambridge, MA, 2023. ISBN 9780262047593. Hardcover. eBook ISBN: 9780262372824.

Macar, U., Bogdan, P. C., Rajamanoharan, S., and Nanda, N. Thought branches: Interpreting llm reasoning requires resampling. *arXiv preprint arXiv:2510.27484*, 2025.

Marr, D. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

McFadden, D. Conditional logit analysis of qualitative choice behavior. In Zarembka, P. (ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press, New York, 1974.

Mialon, G., Fourrier, C., Wolf, T., LeCun, Y., and Scialom, T. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2023.

Moeini, A., Wang, J., Beck, J., Blaser, E., Whiteson, S., Chandra, R., and Zhang, S. A survey of in-context reinforcement learning. *arXiv preprint arXiv:2502.07978*, 2025.

Momentè, F., Suglia, A., Giulianelli, M., Ferrari, A., Koller, A., Lemon, O., Schlangen, D., Fernández, R., and Bernardi, R. Triangulating LLM progress through benchmarks, games, and cognitive tests. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20051–20072, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1092. URL https://aclanthology.org/2025.findings-emnlp.1092/.

Monea, G., Bosselut, A., Brantley, K., and Artzi, Y. LLMs are in-context bandit reinforcement learners. In *Conference on Language Modeling (COLM)*, 2025.

Moskvichev, A., Odouard, V. V., and Mitchell, M. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. *arXiv preprint arXiv:2305.07141*, 2023.

Najemnik, J. and Geisler, W. S. Optimal eye movement strategies in visual search. *Nature*, 434:387–391, 2005. doi: 10.1038/nature03390.

Nie, A., Su, Y., Chang, B., Lee, J. N., Chi, E. H., Le, Q. V., and Chen, M. EVOLVE: Evaluating and optimizing LLMs for exploration. *arXiv preprint arXiv:2410.06238*, 2024.

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv:2112.00114*, 2021.

OpenAI. Gpt-4.1 mini — api model documentation, 2025a. Accessed 2025-09-23.

OpenAI. Gpt-5 system card. Technical report, OpenAI, 2025b. Accessed 2025-09-23.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv*, 2025c.

Ortega, P. A. and Braun, D. A. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469(2153):20120683, May 2013. doi: 10.1098/rspa.2012.0683.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Panwar, M., Ahuja, K., and Goyal, N. In-context learning through the Bayesian prism. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=HX5ujdsSon.

Park, C., Liu, X., Ozdaglar, A., and Zhang, K. Do LLM agents have regret? a case study in online learning and games. *arXiv preprint arXiv:2403.16843*, 2024.

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., and Griffiths, T. L. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214, 2021.

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Qwen Team. Qwen3-235b (moe) — hugging face model card, 2025. Accessed 2025-09-23.

Rahn, N., Bellemare, M. G., and D'Oro, P. Controlling large language model agents with entropic activation steering. *arXiv preprint arXiv:2406.00244*, 2024.

Ravichander, A., Belinkov, Y., and Hovy, E. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, 2021.

Reynolds, L. and McDonell, K. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pp. 1–7, 2021.

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, 117(4):1144–1167, 2010. doi: 10.1037/a0020511.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

Schubert, J. A., Jagadish, A. K., Binz, M., and Schulz, E. In-context learning agents are asymmetric belief updaters. *arXiv preprint arXiv:2402.03969*, 2024.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Shu, M., Wang, J., Zhu, C., Geiping, J., Xiao, C., and Goldstein, T. On the exploitability of instruction tuning. *Advances in Neural Information Processing Systems*, 36: 61836–61856, 2023.

Slivkins, A. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Sprague, Z., Ye, X., Bostrom, K., Chaudhuri, S., and Durrett, G. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv preprint arXiv:2310.16049*, 2023.

Su, E., Ho, M. K., and Gureckis, T. M. Integration of language and experience via the instructed bandit task. In *Proceedings of the 47th Annual Meeting of the Cognitive Science Society*, volume 47. Cognitive Science Society, 2025. URL https://escholarship.org/uc/item/1j746305. Conference paper.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Turpin, M., Michael, J., Perez, E., and Bowman, S. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36: 74952–74965, 2023.

Usher, M. and McClelland, J. L. The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, 108(3):550–592, 2001. doi: 10.1037/0033-295X.108.3.550.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022a.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022b.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wei, S.-L., Wu, C.-K., Huang, H.-H., and Chen, H.-H. Unveiling selection biases: Exploring order and token sensitivity in large language models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5598–5621, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.333.

xAI. Grok 3 beta — the age of reasoning agents, 2025. Accessed 2025-09-23.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Yang, A. et al. Qwen3 technical report. *arXiv*, 2025.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., and Cao, Y. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

Zeng, A., Lv, X., Zheng, Q., Hou, Z., Chen, B., Xie, C., Wang, C., Yin, D., Zeng, H., Zhang, J., et al. Glm-4.5:

Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Zhang, Y., Zhang, F., Yang, Z., and Wang, Z. What and how does in-context learning learn? Bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*, 2023. URL https://arxiv.org/abs/2305.19420.

Zhao, C., Habule, M., and Zhang, W. Large language models (llms) as research subjects: Status, opportunities and challenges. *New Ideas in Psychology*, 79:101167, 2025.

Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

Contents

A Limitations and Future Work **16**

B Full Benchmark Details **17**
B.1 Prompting Structure used for the LLMs in the Task . . . 17
B.2 Language models used . . . 17
B.3 Sampling loss significance table . . . 18
B.4 Inference loss significance table . . . 19
B.5 Benchmark figure segmented by reasoning regimes and model architectures . . . 20
B.6 Model size, architectures and training regime correlations table . . . 21
B.7 Reasoning traces: avg. number of tokens per round . . . 22
B.8 Reasoning traces for the same game under low- vs. high-reasoning . . . 23
B.9 Exact dynamic-programming ceiling . . . 26

C Reinforcement Learning Agent Details **27**

D Compute Resources Used **27**

E Task Performance across Rounds **28**

F Human Participants **29**

G Behavioral Model reproduces Behavioral Metrics **30**

H Per model parameter values **31**

I Activation extraction and linear probing **32**

J PCA geometry of layer-1 activations **33**

K Model Formulation Details **35**
K.1 Task likelihood and Bayesian reference . . . 35
K.2 Centered evidence and gamma memory . . . 35
K.3 Internal posterior and policy . . . 36
K.4 Model loss function . . . 37
K.5 Fitting Results . . . 38
K.6 Statistical significance of behavioral model fits . . . 44
K.7 Parameter recovery and Parameter dispersion over fits . . . 46
K.8 Ablation experiments: Structured Parameter Sweeps . . . 49
K.9 Model Fitting Parameters . . . 49
K.10 Model-search and selected parametrization . . . 51

A. Limitations and Future Work

Our task and behavioral model provide a step toward evaluations that move beyond final accuracy, which here allowed us to disentangle *sampling* from *inference* and to map both humans and LLMs into a shared space of interpretable latent variables. This framing yields a compact, process-level account of behavior and makes concrete, testable predictions. At the same time, some modeling and benchmarking choices trade expressivity for interpretability, suggesting clear directions for improvement.

(1) Human-LLM interface asymmetry A caveat in comparing invalid-choice rates is that humans interacted with a graphical interface in which unavailable actions could not be selected, whereas LLMs interacted through a text interface and could produce occluded, malformed, or missing responses. We therefore treat invalid-choice rates as an interface-sensitive diagnostic of task compliance, rather than as a direct cognitive comparison between humans and LLMs. Our main human–LLM comparisons are based on success, sampling loss, inference loss, and fitted behavioral parameters, with invalid choices reported separately.

(2) A simple valid-observation sampling schedule. The fitted model sets $\kappa_s(t) = m\frac{n_t}{T}$, so sampling becomes more evidence-guided as valid observations accumulate. This captures a basic within-game transition from exploration to exploitation, but still imposes one monotone slope m shared across all contexts. It cannot capture richer adaptations such as sudden strategy switches, state-dependent exploration after invalid actions, or separate sampling policies for early ambiguous evidence versus late confirmatory evidence. Even where sampling posterior trajectories are imperfectly fit, they provide direct support for the core conclusion that sampling and inference dissociate, and that extended reasoning primarily improves inference while leaving a systematic gap in active information acquisition.

(3) A single set of task meta-parameters. Because collecting statistically reliable human data and running large-scale evaluations across a broad set of LLMs is expensive and time-consuming, we focus on a single regime of task meta-parameters (number of buttons, emission probabilities, occlusion process, and horizon distribution). Different regimes may induce qualitatively different optimal strategies and could change both the magnitude and the nature of *reasoning* effects, motivating systematic sweeps over these task parameters.

Overall, our framework points toward a next generation of evaluations: benchmarks that measure capability while constraining the computations that produce it, and behavioral models that translate behavior into algorithmic variables, setting the stage for implementation-level tests of how reasoning is realized in language-model circuits.

B. Full Benchmark Details

B.1. Prompting Structure used for the LLMs in the Task

We evaluate language models using a text-based version of the task that mirrors the instructions and feedback given to human participants. The prompt specifies the task mechanics (four cues, one biased at 90/10 RED/GREEN while the others are 50/50), the round-by-round interaction protocol (available cues may change across rounds, but at least one cue is always available), and a strict response format (one letter per round, no additional text). After the sampling rounds, models are asked to report which cue had the highest proportion of RED outcomes (the biased cue). The prompt also defines the feedback and scoring scheme (correct/incorrect/invalid; +100 for a correct final report and -100 otherwise), explicitly framing success as identifying the biased cue. Models interact with the environment across full trials and receive round- and trial-specific updates conditional on their previous choices.

Active Probabilistic Reasoning Task Prompt

```
Task
- You will play a game with \{n\_rounds\} rounds.
- In each round, some cues are available: A, B, C, D.
- One cue is biased: 90\% RED / 10\% GREEN. The others are 50/50.
- Available cues may disappear at random, but at least one is always active.
- Each round, respond with exactly one letter (A, B, C, or D). No markup or punctuation.

- After \{n\_rounds\} rounds, identify the biased cue.
- Scoring: Correct +100 points, Wrong -100 points.

Round prompt
- ``Trial X, Round \{current\_round\}: Available cues \{available\_cues\}. Which do you
     choose? Respond with exactly one letter: A, B, C, or D.''

Round feedback
- ``Trial X, Round \{current\_round\}: Available cues \{available\_cues\}. You chose ``
    CURRENT\_ANSWER'' and saw ``RESULT''.''

Final decision prompt
- ``Trial X: Based on all observed colors, which cue \{letters\} was biased toward red?
     Respond with exactly one letter: A, B, C, or D.''

Final feedback
- ``Trial X: Based on all observed colors, which cue \{letters\} was biased toward red?
   You chose ``CURRENT\_ANSWER'' which was ``FEEDBACK-CORRECT/INCORRECT''.
   You received \{score\} points.''

Feedback labels
- Correct: ``the biased cue''
- Incorrect: ``not the biased cue. The biased cue was ``BIASED\_QUADRANT''.''
- Invalid: ``an invalid choice''
```

B.2. Language models used

The full set of models benchmarked in this task is: OpenAI's *gpt 4o mini* (Hurst et al., 2024), *gpt 4.1 mini* (OpenAI, 2025a), *gpt 5 mini* (OpenAI, 2025b), and the *gpt oss* open architectures in both the 20b and 120b parameter variants (OpenAI, 2025c); several *llama* models (Touvron et al., 2023; Dubey et al., 2024), including variants fine-tuned on human behavioral data (Binz et al., 2025); a distilled version of *deepseek* (Guo et al., 2025b); Anthropic's *claude sonnet 4* and *claude haiku 3.5* (Anthropic, 2025); Google's *gemini 2.5 pro/flash* and the smaller *gemma* models (Comanici et al., 2025; Gemma Team, Google DeepMind, 2025); the *qwen* family, including the 235B Mixture-of-Experts (MoE) model and earlier dense variants (Qwen Team, 2025; Yang et al., 2025); and the fully open-source *apertus* model (Hernández-Cano et al., 2025), *grok 3 mini* (xAI, 2025), and *glm 4.5* (Zeng et al., 2025).

B.3. Sampling loss significance table

Table 1. Sampling loss comparison between human top-25% performance and all tested models (lower values indicate better sampling). Human top-25% mean sampling loss = 0.0080. Bonferroni correction applied across 50 tests.

Model	Reasoning	N	LLM loss	Human loss	Diff	95% CI	p	p_{Bonf}	Sig.
gpt-5-mini	base	5800	0.138	0.008	-0.130	[-0.159, -0.102]	$< 10^{-4}$	$< 10^{-4}$	***
grok-3-mini	base	6200	0.121	0.008	-0.113	[-0.142, -0.084]	$< 10^{-4}$	$< 10^{-4}$	***
gemini-2.5-flash	base	3000	0.135	0.008	-0.127	[-0.158, -0.096]	$< 10^{-4}$	$< 10^{-4}$	***
gemini-2.5-pro	base	3000	0.136	0.008	-0.128	[-0.158, -0.096]	$< 10^{-4}$	$< 10^{-4}$	***
gpt-oss-120b	base	3000	0.137	0.008	-0.129	[-0.159, -0.098]	$< 10^{-4}$	$< 10^{-4}$	***
gpt-5	base	3000	0.133	0.008	-0.125	[-0.156, -0.094]	$< 10^{-4}$	$< 10^{-4}$	***
gpt-oss-20b	base	3000	0.170	0.008	-0.162	[-0.193, -0.130]	$< 10^{-4}$	$< 10^{-4}$	***
claude-sonnet-4	base	5800	0.109	0.008	-0.101	[-0.130, -0.072]	$< 10^{-4}$	$< 10^{-4}$	***
glm-4.5	base	3200	0.163	0.008	-0.155	[-0.186, -0.124]	$< 10^{-4}$	$< 10^{-4}$	***
DeepSeek-R1-Qwen3-8B	base	3000	0.125	0.008	-0.117	[-0.149, -0.086]	$< 10^{-4}$	$< 10^{-4}$	***
gpt4.1-mini	base	4500	0.122	0.008	-0.114	[-0.144, -0.085]	$< 10^{-4}$	$< 10^{-4}$	***
claude-haiku-3.5	base	3000	0.145	0.008	-0.137	[-0.168, -0.106]	$< 10^{-4}$	$< 10^{-4}$	***
Qwen3_235B_Instruct	base	1405	0.121	0.008	-0.114	[-0.149, -0.077]	$< 10^{-4}$	$< 10^{-4}$	***
gpt4o-mini	base	1500	0.130	0.008	-0.122	[-0.158, -0.086]	$< 10^{-4}$	$< 10^{-4}$	***
Qwen_14B_Instruct	base	1560	0.130	0.008	-0.122	[-0.157, -0.087]	$< 10^{-4}$	$< 10^{-4}$	***
Llama-4-Scout-Instruct	base	3000	0.146	0.008	-0.138	[-0.169, -0.107]	$< 10^{-4}$	$< 10^{-4}$	***
Qwen_7B_Instruct	base	1500	0.112	0.008	-0.104	[-0.139, -0.069]	$< 10^{-4}$	$< 10^{-4}$	***
Llama-3.1-405B-Instruct	base	3000	0.169	0.008	-0.161	[-0.192, -0.129]	$< 10^{-4}$	$< 10^{-4}$	***
Llama-4-Maverick-Instruct	base	3000	0.160	0.008	-0.152	[-0.183, -0.121]	$< 10^{-4}$	$< 10^{-4}$	***
Llama_3_3_70B_Instruct	base	3000	0.149	0.008	-0.141	[-0.172, -0.110]	$< 10^{-4}$	$< 10^{-4}$	***
Gemma_2B_Instruct	base	1540	0.130	0.008	-0.122	[-0.158, -0.086]	$< 10^{-4}$	$< 10^{-4}$	***
Centaur_70B	base	1600	0.137	0.008	-0.129	[-0.165, -0.093]	$< 10^{-4}$	$< 10^{-4}$	***
Centaur_8B	base	1500	0.148	0.008	-0.140	[-0.175, -0.103]	$< 10^{-4}$	$< 10^{-4}$	***
Apertus_70B_Instruct	base	2300	0.164	0.008	-0.156	[-0.189, -0.123]	$< 10^{-4}$	$< 10^{-4}$	***
Qwen3_32B	base	1500	0.137	0.008	-0.129	[-0.165, -0.093]	$< 10^{-4}$	$< 10^{-4}$	***
Apertus_8B_Instruct	base	2900	0.185	0.008	-0.177	[-0.209, -0.146]	$< 10^{-4}$	$< 10^{-4}$	***
Qwen_3B_Instruct	base	1540	0.156	0.008	-0.148	[-0.184, -0.112]	$< 10^{-4}$	$< 10^{-4}$	***
Llama-3.2-3B-Instruct	base	1500	0.142	0.008	-0.134	[-0.170, -0.099]	$< 10^{-4}$	$< 10^{-4}$	***
Extended reasoning models									
gpt-5-mini	extended	3000	0.102	0.008	-0.094	[-0.124, -0.063]	$< 10^{-4}$	$< 10^{-4}$	***
grok-3-mini	extended	3200	0.107	0.008	-0.099	[-0.129, -0.068]	$< 10^{-4}$	$< 10^{-4}$	***
gemini-2.5-flash	extended	3000	0.104	0.008	-0.096	[-0.127, -0.065]	$< 10^{-4}$	$< 10^{-4}$	***
gemini-2.5-pro	extended	3000	0.088	0.008	-0.080	[-0.111, -0.049]	$< 10^{-4}$	$< 10^{-4}$	***
gpt-oss-120b	extended	3000	0.107	0.008	-0.099	[-0.130, -0.068]	$< 10^{-4}$	$< 10^{-4}$	***
gpt-5	extended	3000	0.121	0.008	-0.113	[-0.144, -0.082]	$< 10^{-4}$	$< 10^{-4}$	***
gpt-oss-20b	extended	6000	0.120	0.008	-0.112	[-0.140, -0.084]	$< 10^{-4}$	$< 10^{-4}$	***
glm-4.5	extended	3000	0.138	0.008	-0.130	[-0.160, -0.098]	$< 10^{-4}$	$< 10^{-4}$	***

B.4. Inference loss significance table

Table 2. Inference loss comparison between human top-25% performance and all tested models (lower values indicate better inference). Human top-25% mean inference loss = 0.0150. Bonferroni correction applied across 50 tests.

Model	Reasoning	N	LLM loss	Human loss	Diff	95% CI	p	p_{Bonf}	Sig.
gpt-5-mini	base	5800	0.111	0.015	-0.096	[-0.117, -0.074]	$< 10^{-4}$	$< 10^{-4}$	***
grok-3-mini	base	6200	0.017	0.015	-0.002	[-0.022, +0.019]	0.9201	1.0000	n.s.
gemini-2.5-flash	base	3000	0.220	0.015	-0.205	[-0.231, -0.178]	$< 10^{-4}$	$< 10^{-4}$	***
gemini-2.5-pro	base	3000	0.173	0.015	-0.158	[-0.184, -0.132]	$< 10^{-4}$	$< 10^{-4}$	***
gpt-oss-120b	base	3000	0.084	0.015	-0.069	[-0.092, -0.045]	$< 10^{-4}$	$< 10^{-4}$	***
gpt-5	base	3000	0.147	0.015	-0.132	[-0.158, -0.107]	$< 10^{-4}$	$< 10^{-4}$	***
gpt-oss-20b	base	3000	0.100	0.015	-0.085	[-0.109, -0.061]	$< 10^{-4}$	$< 10^{-4}$	***
claude-sonnet-4	base	5800	0.032	0.015	-0.017	[-0.037, +0.003]	0.1178	1.0000	n.s.
glm-4.5	base	3200	0.194	0.015	-0.179	[-0.205, -0.153]	$< 10^{-4}$	$< 10^{-4}$	***
DeepSeek-R1-Qwen3-8B	base	3000	0.082	0.015	-0.067	[-0.091, -0.043]	$< 10^{-4}$	$< 10^{-4}$	***
gpt4.1-mini	base	4500	0.111	0.015	-0.096	[-0.120, -0.073]	$< 10^{-4}$	$< 10^{-4}$	***
claude-haiku-3.5	base	3000	0.142	0.015	-0.127	[-0.153, -0.102]	$< 10^{-4}$	$< 10^{-4}$	***
Qwen3_235B_Instruct	base	1405	0.204	0.015	-0.189	[-0.223, -0.156]	$< 10^{-4}$	$< 10^{-4}$	***
gpt4o-mini	base	1500	0.221	0.015	-0.206	[-0.238, -0.173]	$< 10^{-4}$	$< 10^{-4}$	***
Qwen_14B_Instruct	base	1560	0.221	0.015	-0.206	[-0.239, -0.173]	$< 10^{-4}$	$< 10^{-4}$	***
Llama-4-Scout-Instruct	base	3000	0.216	0.015	-0.201	[-0.228, -0.175]	$< 10^{-4}$	$< 10^{-4}$	***
Qwen_7B_Instruct	base	1500	0.275	0.015	-0.260	[-0.295, -0.225]	$< 10^{-4}$	$< 10^{-4}$	***
Llama-3.1-405B-Instruct	base	3000	0.225	0.015	-0.210	[-0.237, -0.183]	$< 10^{-4}$	$< 10^{-4}$	***
Llama-4-Maverick-Instruct	base	3000	0.252	0.015	-0.237	[-0.263, -0.211]	$< 10^{-4}$	$< 10^{-4}$	***
Llama_3_3_70B_Instruct	base	3000	0.263	0.015	-0.248	[-0.275, -0.221]	$< 10^{-4}$	$< 10^{-4}$	***
Gemma_2B_Instruct	base	1540	0.283	0.015	-0.268	[-0.303, -0.232]	$< 10^{-4}$	$< 10^{-4}$	***
Centaur_70B	base	1600	0.298	0.015	-0.283	[-0.315, -0.250]	$< 10^{-4}$	$< 10^{-4}$	***
Centaur_8B	base	1500	0.301	0.015	-0.286	[-0.320, -0.253]	$< 10^{-4}$	$< 10^{-4}$	***
Apertus_70B_Instruct	base	2300	0.294	0.015	-0.279	[-0.309, -0.248]	$< 10^{-4}$	$< 10^{-4}$	***
Qwen3_32B	base	1500	0.326	0.015	-0.311	[-0.343, -0.278]	$< 10^{-4}$	$< 10^{-4}$	***
Apertus_8B_Instruct	base	2900	0.287	0.015	-0.272	[-0.300, -0.244]	$< 10^{-4}$	$< 10^{-4}$	***
Qwen_3B_Instruct	base	1540	0.320	0.015	-0.305	[-0.339, -0.270]	$< 10^{-4}$	$< 10^{-4}$	***
Llama-3.2-3B-Instruct	base	1500	0.343	0.015	-0.328	[-0.367, -0.292]	$< 10^{-4}$	$< 10^{-4}$	***
Extended reasoning models									
gpt-5-mini	extended	3000	0.013	0.015	+0.002	[-0.018, +0.022]	0.8430	1.0000	n.s.
grok-3-mini	extended	3200	0.019	0.015	-0.004	[-0.025, +0.017]	0.7234	1.0000	n.s.
gemini-2.5-flash	extended	3000	0.021	0.015	-0.006	[-0.027, +0.014]	0.5332	1.0000	n.s.
gemini-2.5-pro	extended	3000	0.042	0.015	-0.027	[-0.048, -0.006]	0.0159	0.7975	*
gpt-oss-120b	extended	3000	0.024	0.015	-0.009	[-0.030, +0.013]	0.4517	1.0000	n.s.
gpt-5	extended	3000	0.024	0.015	-0.009	[-0.031, +0.013]	0.4076	1.0000	n.s.
gpt-oss-20b	extended	6000	0.030	0.015	-0.015	[-0.034, +0.005]	0.1498	1.0000	n.s.
glm-4.5	extended	3000	0.056	0.015	-0.041	[-0.063, -0.019]	0.0007	0.0335	***

B.5. Benchmark figure segmented by reasoning regimes and model architectures

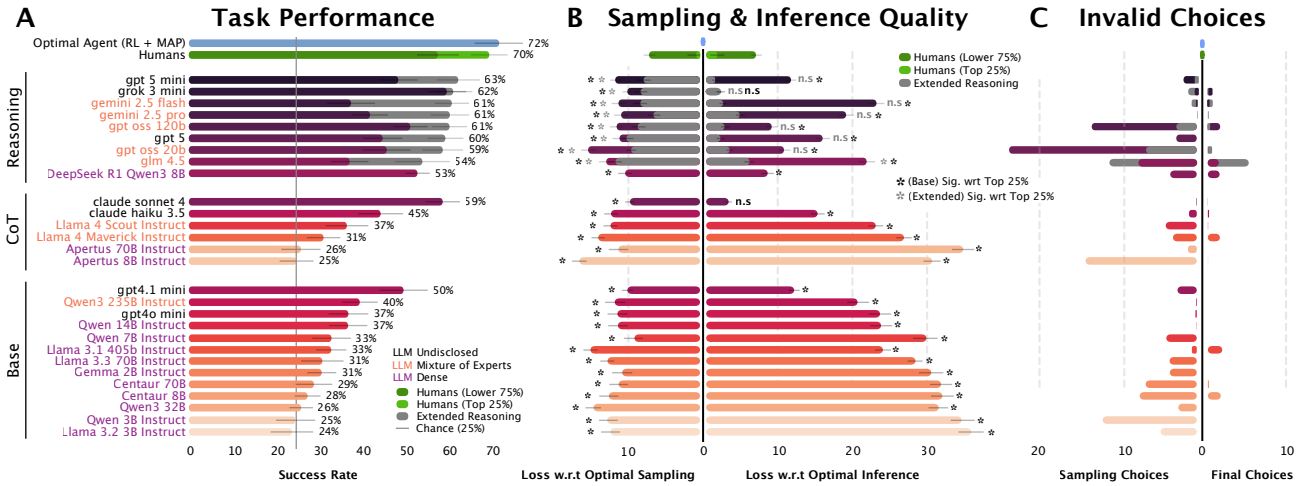

Figure 5. **Comparing human and LLM behavior across reasoning regimes and architectures.** Models are segmented into three groups: *Reasoning* (post-RL models that expose a controllable reasoning-effort parameter, evaluated under both low and high reasoning effort; the high-effort condition is shown as Extended Reasoning with gray bars), *CoT* (models that spontaneously produce a chain-of-thought trace despite instructions to respond with a single token, allowed up to 500 tokens with the final answer extracted), and *Base* (models constrained to a single-token response via `max_tokens=1`, preventing any form of intermediate reasoning). N.B.: `Claude Sonnet 4` supports extended reasoning but was not evaluated in that condition due to API costs, and is therefore placed in the CoT group. `DeepSeek R1 Qwen3 8B` does not allow control of reasoning effort. Model names are color-coded by architecture type: black for undisclosed, orange for mixture-of-experts, and purple for dense. **A: Task performance.** Average success rate across trial lengths $T \in \{2, \ldots, 15\}$. We report human performance (green), split into lower 75% and top 25% of participants, and the optimal agent (PPO sampling + MAP inference) (light blue). For Reasoning models that support increased reasoning effort, gray overlays indicate the Extended Reasoning condition. Error bars represent standard deviations, computed across trial-cluster means with a uniform distribution over the number of rounds. The vertical line marks chance performance (25%). A clear hierarchy emerges: post-RL Reasoning models consistently outperform CoT models, which in turn outperform Base models. **B: Sampling and inference quality.** We quantify sampling quality (left) and inference quality (right) as performance loss with respect to the optimal agent (lower is better): inference loss is the gap between the agent and a counterfactual agent that preserves the same sampled evidence but applies MAP at the inference round for the final decision; sampling loss is the gap between this counterfactual MAP agent and the reference agent (PPO + MAP), isolating suboptimal evidence acquisition. Error bars show bootstrap standard deviations ($B = 1{,}000$, paired resampling at the trial level). Asterisks denote significance of a two-tailed permutation test comparing each model's per-trial loss against the human top-25% baseline ($*$: $p < 0.0001$; n.s.: not significant; Bonferroni-corrected across all comparisons); black asterisks correspond to the base condition and gray ones to extended reasoning. Extended reasoning primarily reduces inference loss, with only modest effects on sampling loss. Extended Reasoning gains (gray bars) are exclusive to the post-RL Reasoning group, as only these models expose a controllable reasoning-effort parameter. **C: Invalid choices.** Fraction of invalid choices during sampling (left) and at the final inference decision (right). Invalid choices include selecting occluded options, producing tokens outside the valid choice set (A–D), or failing to respond; humans cannot produce invalid choices in the graphical interface. Invalid choices occur more frequently during sampling than at the final decision and are reduced by reasoning.

B.6. Model size, architectures and training regime correlations table

Table 3. **Training regime, not model scale, is the dominant predictor of task performance and cognitive-space coordinates.** R^2 values from linear regressions predicting each behavioral parameter from model covariates, computed across 19 open-source models for which model size, architecture (dense vs. MoE), and training regime (Base, CoT, Reasoning) are known. Training regime alone explains the most variance in both task success ($R^2 = 0.862$) and the inference strategy parameter κ_f ($R^2 = 0.589$), substantially outperforming raw model scale ($R^2 = 0.088$ and 0.147, respectively). Architecture is also informative, particularly for success and κ_f, and combining training regime with architecture yields strong explanatory power ($R^2 = 0.906$ for success; 0.756 for κ_f). Adding model size on top provides negligible further lift ($0.906 \rightarrow 0.908$; $0.756 \rightarrow 0.757$), confirming that *how* a model is trained, rather than how large it is, primarily determines its location in cognitive space. In contrast, the memory parameter β is poorly explained by any combination of these factors ($R^2 \leq 0.217$), suggesting it captures individual model idiosyncrasies, while bias spread is best captured by the combination of size and regime ($R^2 = 0.433$).

Predictors	Success rate	β	κ_f	m	Bias spread
Single and pairwise predictors					
Model size	0.088	0.116	0.147	0.059	0.084
Architecture	0.416	0.002	0.555	0.172	0.004
Training regime	0.862	0.068	0.589	0.305	0.319
Model size + architecture	0.428	0.154	0.561	0.172	0.173
Training regime + architecture	0.906	0.105	0.756	0.338	0.348
Full predictor sets					
Model size + training regime	0.889	0.214	0.666	0.333	0.433
Model size + train. regime + arch.	0.908	0.217	0.757	0.343	0.435

B.7. Reasoning traces: avg. number of tokens per round

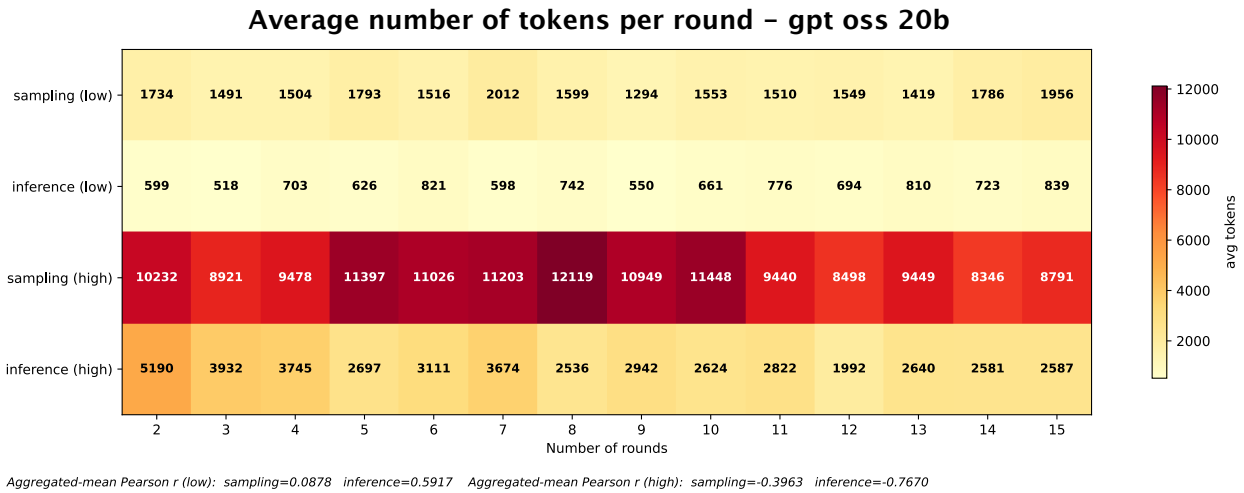

Figure 6. **Reasoning token budget does not explain the selective improvement in inference.** Average number of reasoning tokens produced by gpt oss 20b per round, shown separately for sampling and inference phases under low- and high-reasoning conditions, across all game lengths ($T \in \{2, \ldots, 15\}$). Under high reasoning, sampling rounds consume substantially more tokens ($\sim$8,000–12,000) than inference decisions ($\sim$2,000–5,000). If extended reasoning improved performance primarily by providing a larger compute budget, the phase receiving more tokens should benefit most. Instead, the opposite holds: reasoning reduces inference loss by 10.9 pp but sampling loss by only 3.0 pp. This dissociation is difficult to reconcile with a pure compute-budget explanation, and instead suggests that reasoning selectively enables the kind of computation relevant to evidence integration—consistent with the behavioral model's account of improved inference via more accurate memory ($\beta \approx 0$) and sharper belief-to-choice mapping (larger κ_f).

B.8. Reasoning traces for the same game under low- vs. high-reasoning

Task Prompt and Evidence History — Correct final answer: A

```
Task: You will play a game with 15 rounds. In each round you'll see available cues (
    choosable):

One cue has 90\% RED / 10\% GREEN, others have 50/50 color distribution. Possible cues
    are: A, B, C, D. Available cues disappear after random duration, at least one cue
    is active per round. Your task is to pick one of the available cues every round, by
     responding with just the letter and nothing else. Don't use markup or punctuation.
     After 15 rounds, identify the cue biased towards RED. Correct: +100 points, Wrong:
     -100 points.

Trial 0, Round 1: Available cues B, C, D. You chose "C" and saw RED.
Trial 0, Round 2: Available cues C. You chose "C" and saw RED.
Trial 0, Round 3: Available cues A, B, C, D. You chose "B" and saw GREEN.
Trial 0, Round 4: Available cues D. You chose "D" and saw GREEN.
Trial 0, Round 5: Available cues C. You chose "C" and saw RED.
Trial 0, Round 6: Available cues C. You chose "C" and saw RED.
Trial 0, Round 7: Available cues A, B, C. You chose "A" and saw RED.
Trial 0, Round 8: Available cues C, D. You chose "C" and saw GREEN.
Trial 0, Round 9: Available cues C, D. You chose "D" and saw RED.
Trial 0, Round 10: Available cues A, B, C, D. You chose "C" and saw RED.
Trial 0, Round 11: Available cues B, C. You chose "C" and saw GREEN.
Trial 0, Round 12: Available cues A, D. You chose "A" and saw RED.
Trial 0, Round 13: Available cues B. You chose "B" and saw RED.
Trial 0, Round 14: Available cues A, C, D. You chose "A" and saw RED.
Trial 0, Round 15: Available cues A, B. You chose "A" and saw GREEN.

Trial 0: Based on all observed colors, which cue A, B, C, or D do you think was biased
    toward red? Respond with exactly one letter: A, B, C, or D.
```

Low-Reasoning trace (gpt-oss-20b) — Final answer: C

```
We saw C twice both RED. So C likely biased. Return C.
```

High-Reasoning Trace (gpt-oss-20b) — Final answer: A

```
We selected various cues each round. We saw a sequence of colors. We need to determine
    which of A, B, C, D is the biased one. We are to respond with a letter.
Thus it's a statistical inference problem: From the data, find which cue has the
    highest fraction of RED observed vs total times chosen. But we also must consider
    that some cues might not have been chosen if they were not available? In our logs
    we always chose a cue. We might not have chosen many times for a certain cue; but
    we need to see which gave the highest RED ratio across all times we chose that cue.
     However, the data may show biases due to chance. We need to compute the frequency
    of RED for each cue for the times we chose that cue.
However, note that we might not have chosen A many times, but the ratio could still be
    high. But does the question ask for "biased towards red"? Yes: "Based on all
    observed colors, which cue A, B, C, or D do you think was biased toward red?" So we
     just compute ratio of RED across our sample of times selecting each cue. This
    might not reflect the actual bias of each cue, because we only choose each cue a
    few times. But maybe the ratio among our selected times reveals the actual bias.
    Let's calculate.

List each round, chosen cue, color:

Round 1: C, RED
```

```
Round 2: C, RED
Round 3: B, GREEN
Round 4: D, GREEN
Round 5: C, RED
Round 6: C, RED
Round 7: A, RED
Round 8: C, GREEN
Round 9: D, RED
Round 10: C, RED
Round 11: C, GREEN
Round 12: A, RED
Round 13: B, RED
Round 14: A, RED
Round 15: A, GREEN

Now count for each cue.

Cue A: Rounds 7, 12, 14, 15. That's 4 times. Colors: 7: RED, 12: RED, 14: RED, 15:
    GREEN. So RED count = 3, GREEN count = 1. Ratio RED = 3/4 = 0.75.
Cue B: Rounds 3 and 13. 2 times. Colors: 3: GREEN, 13: RED. So 1 RED, 1 GREEN. Ratio
    RED = 0.5.
Cue C: Rounds 1,2,5,6,8,10,11. That's 7 times. 1: RED, 2: RED, 5: RED, 6: RED, 8: GREEN
    , 10: RED, 11: GREEN. So RED count = 5, GREEN count = 2. Ratio RED = 5/7 approx.
    0.7142857.
Cue D: Rounds 4 and 9. 2 times. Colors: 4: GREEN, 9: RED. So 1 RED, 1 GREEN. Ratio RED
    = 0.5.

So the highest ratio is for cue A: 0.75. Cue C: 0.714... Cue B and D: 0.5.

But we only have 4 samples for A. There's potential for small sample variation. But the
     question is to identify based on observed colors the cue with highest ratio.
    According to our data, it's A. But we need to determine if any other had a higher
    ratio? No. So answer: A.
We should output exactly one letter: A, B, C, or D. So the correct letter is A.

But let's double-check if any nuance: One cue is 90\% of one color, 10\% of other.
    Suppose the biased cue is not necessarily the one biased toward red. In fact, we
    are not told which color is the 90\% color. It could be RED or GREEN. We only know
    that one cue is 90\% one color, 10\% the other; the others are 50/50 split. So we
    need to identify which cue is biased. How to determine that? Many possibilities.
The simplest method: Use the sample data as evidence. We want to find the cue with the
    largest difference between observed proportions and 0.5. But the problem states: "
    Based on all observed colors, which cue A, B, C, or D do you think was biased
    toward red?" So we are not asked to identify which cue is biased overall; we are
    asked to pick the one biased toward red. So we answer A.
But could there be reasoning that the biased cue is the one biased toward red? Possibly
     the biased cue's "color" is unknown: maybe the 90\% color could be RED or GREEN.
    We can only compute RED ratio. But the question explicitly asks which cue was
    biased toward red. So answer that.
Thus answer = A. Sure.

But let's consider if the 90\% color is GREEN for a cue. Then its RED ratio would be
    10\%. That might appear low. So the highest RED ratio would then be a 50/50 cue?
    Actually, if the biased cue's high color is GREEN, then its RED ratio would be
    lower than the unbiased ones. So picking the cue with highest RED ratio might not
    identify the biased cue. But the instructions say we need to identify the biased
    cue for the bonus points. But the final question is to answer which cue was biased
    toward red. So presumably that is the way to pick the biased cue.
Wait, but the description: "Your task is to pick one of the available cues every round,
     by responding with just the letter and nothing else. ... After 15 rounds, identify
     the biased cue. Correct: +100 points, Wrong: -100 points."
So we must identify the biased cue. The question at the end says, "Based on all
    observed colors, which cue A, B, C, or D do you think was biased toward red?" This
```

```
    is presumably the answer. But the question to us is: answer with exactly one letter:
     A, B, C, or D. So we answer A.
Thus output: "A".

But check careful: Did we miscount? Let's recount.

List all chosen clues:
Round 1: C, RED
Round 2: C, RED
Round 3: B, GREEN
Round 4: D, GREEN
Round 5: C, RED
Round 6: C, RED
Round 7: A, RED
Round 8: C, GREEN
Round 9: D, RED
Round 10: C, RED
Round 11: C, GREEN
Round 12: A, RED
Round 13: B, RED
Round 14: A, RED
Round 15: A, GREEN

Count:
C appears 1,2,5,6,8,10,11 -> 7 times. Of these, RED at 1,2,5,6,10 = 5 times. GREEN at
    8,11 = 2 times. Ratio = 5/7 ~ 0.7142857.
A appears at 7,12,14,15 -> 4 times. RED at 7,12,14 -> 3 times. GREEN at 15 -> 1 time.
    Ratio = 3/4 = 0.75.
B appears at 3,13 -> 2 times. GREEN at 3, RED at 13 -> ratio = 1/2 = 0.5.
D appears at 4,9 -> 2 times. GREEN at 4, RED at 9 -> ratio = 1/2 = 0.5.

Thus A was biased toward red. So answer A.

But we might want to consider the possibility that the 90\% color could be GREEN. In
    that case, the biased cue's RED ratio would be 10\%. So the highest RED ratio would
     be between unbiased 50/50 cues and maybe one unbiased might happen to have more
    RED than 50\% due to randomness. Here we found A has 75\% RED, which is quite high
    relative to 50\%. Unbiased could by chance also show 75\% RED if e.g. the sample
    draws were not large. But we have 4 picks for A, and 75\% chance of 75\% for a
    random 50/50 distribution is plausible? The probability of a binomial(4,0.5) having
     at least 3 successes (RED) is C(4,3)*0.5^4 + C(4,4)*0.5^4 = (4+1)/16 = 5/16 $\
    approx$ 0.3125. So not too improbable. For C, 5/7 ratio: probability 5 or more
    successes out of 7 for 50/50 distribution is like? Let's compute: probability of 5
    successes: C(7,5)*0.5^7 = 21/128 $\approx$ 0.16406; 6 successes: C(7,6)*0.5^7 =
    7/128 $\approx$ 0.0546875; 7 successes: 1/128 $\approx$ 0.0078125; sum = ~0.2775.
    So around 27.75\% chance.
Thus these counts could be random occurrences. But maybe the biased cue has 90\%
    probability for one color. If that color is RED, we'd expect a huge number of RED
    in the sample. But we only see 10 RED out of 15 choices. Actually, we saw 10 RED, 5
     GREEN across all 15 rounds. If one cue had 90\% RED, but we only selected cues at
    random, the expected overall count might be somewhere between 0.5 and 0.9 times the
     number of times we selected that cue. If we only selected that cue rarely, maybe
    we wouldn't see many RED. We selected C more frequently (7 times). But maybe C is
    the biased cue? 5 of 7 are RED. That's 71.4\% RED, which is higher than 50\% but
    lower than 90\%. For A, 3 of 4 are RED. That's 75\%. But again, not near 90\%. For
    D, 1/2. For B, 1/2. So perhaps the biased cue is A or C. But which one to pick? The
     question:
"Based on all observed colors, which cue A, B, C, or D do you think was biased toward
    red?\" That is A. So we answer A.
However, might they want to identify the biased cue, not just highest ratio? The
    instructions: "Based on all observed colors, which cue A, B, C, or D do you think
    was biased toward red?" The user answer is a guess: we just need to output that
    letter. So we will output A.
```

```
Thus the answer is "A".
```

B.9. Exact dynamic-programming ceiling

The sampling problem can be solved exactly because the task is a finite-horizon belief-state Markov decision process. Let $k \in \{1, \ldots, K\}$ denote the latent biased button, with uniform prior $p_0(k) = 1/K$. At each sampling round t, the agent observes an availability set $A_t \subseteq \{1, \ldots, K\}$, chooses a visible button $a_t \in A_t$, and observes an outcome $x_t \in \{R, G\}$. The likelihood is

$$P(R \mid k, a) = \alpha_U + (\alpha_B - \alpha_U)\mathbf{1}\{a = k\}, \qquad P(G \mid k, a) = 1 - P(R \mid k, a). \tag{8}$$

The full history can be summarized without loss by the count state

$$c_t = \{n_{i,R}(t), n_{i,G}(t)\}_{i=1}^{K}. \tag{9}$$

The posterior induced by c_t is

$$p_c(k) = \frac{\prod_{i=1}^{K} P(R \mid k, i)^{n_{i,R}} P(G \mid k, i)^{n_{i,G}}}{\sum_{k'=1}^{K} \prod_{i=1}^{K} P(R \mid k', i)^{n_{i,R}} P(G \mid k', i)^{n_{i,G}}}. \tag{10}$$

Thus, conditional on c_t and the current availability set A_t, the distribution of future observations is independent of the earlier history.

Let $V_h(c)$ be the optimal expected final accuracy with h sampling rounds remaining. With no samples remaining, the optimal final decision is MAP:

$$V_0(c) = \max_k p_c(k). \tag{11}$$

For $h \geq 1$, define the predictive probability of RED after sampling button a:

$$\rho(c, a) = P(R \mid c, a) = \sum_k p_c(k) P(R \mid k, a). \tag{12}$$

The action value is

$$Q_h(c, a) = \rho(c, a) V_{h-1}(c + e_{a,R}) + (1 - \rho(c, a)) V_{h-1}(c + e_{a,G}), \tag{13}$$

where $e_{a,R}$ and $e_{a,G}$ increment the corresponding count. Since the occlusion mask is observed before the sampling action, the optimal policy is

$$\pi_h^\star(c, A) = \arg\max_{a \in A} Q_h(c, a). \tag{14}$$

Before observing the next availability set, the value averages over the occlusion process:

$$V_h(c) = \mathbb{E}_A \left[\max_{a \in A} Q_h(c, a) \right]. \tag{15}$$

The exact ceiling for horizon H is therefore

$$J_H^\star = V_H(0). \tag{16}$$

For the task occlusion process, each button is independently visible with probability q, with a fallback that makes one button visible uniformly if no button is visible. Hence, for nonempty A,

$$P(A) = q^{|A|}(1 - q)^{K - |A|} + \mathbf{1}\{|A| = 1\} \frac{(1 - q)^K}{K}. \tag{17}$$

Optimality follows by backward induction. The base case $h = 0$ is optimal because MAP maximizes posterior probability of a correct final decision. Assuming V_{h-1} is optimal for all count states with $h - 1$ samples remaining, any policy with h samples remaining must first choose some legal $a \in A$, whose expected continuation value is $Q_h(c, a)$. Choosing the maximizing action therefore dominates all other first actions, and the induction hypothesis gives the optimal continuation.

For $K = 4$ and $H = 15$, the number of count states is

$$\sum_{t=0}^{15} \binom{t+7}{7} = \binom{23}{8} = 490{,}314, \tag{18}$$

making exact backward induction tractable. With $K = 4$, $\alpha_B = 0.9$, $\alpha_U = 0.5$, and $q = 0.5$, the exact DP ceiling averaged over horizons $T = 2, \ldots, 15$ lies within the performance variance observed in the multi-seed search done with the PPO RL agents (Appendix **C**).

C. Reinforcement Learning Agent Details

We train a reinforcement-learning agent to obtain a near-optimal sampling policy for our task. Training is performed with Proximal Policy Optimization (PPO) (Schulman et al., 2017). At each round, the agent observes a 12-dimensional state vector consisting of (i) cue–color counts for each button (8 dimensions: 4 cues $\times$ 2 colors) and (ii) a 4-dimensional binary mask indicating which cues are currently available. Environment dynamics (cue bias, occlusions, and trial length) match the experimental task, with the number of rounds T sampled uniformly from $\{2, \ldots, 15\}$.

During the sampling phase, the PPO policy selects which cue to sample. At the end of the trial, the agent makes a final decision using a Maximum-A-Posteriori (MAP) rule given the accumulated evidence. Rewards are sparse and tied to task success: the agent receives $+100$ for a correct final report, 0 for valid sampling choices, and -10 for selecting an unavailable cue. Unavailable selections advance the round without revealing evidence, encouraging the agent to avoid invalid choices and sample efficiently.

We select optimization and architecture hyperparameters via a grid search that maximizes correct cue identification. The policy network is a two-layer MLP with 64 hidden units per layer and ReLU activations. We train for 5×10^6 environment steps with learning rate 2×10^{-5} and minibatch size 128. Each PPO iteration collects $T = 2048$ on-policy steps and performs 10 optimization epochs. We use discount $\gamma = 0.99$, generalized advantage estimation $\lambda_{\text{GAE}} = 0.95$, and clipping parameter $\varepsilon = 0.2$, with entropy and value loss coefficients set to 0.01 and 0.5, respectively. Reported performance on the *Active Probabilistic Reasoning Task* was $70\% \pm 2\%$

D. Compute Resources Used

Locally executed experiments were run on a shared GPU server equipped with four NVIDIA H200 GPUs. GPU workloads included PPO training for the near-optimal sampling reference, local inference for open-weight models that fit on the server, and JAX-based batched fitting of the behavioral model. The PPO agent was trained for 5×10^6 environment steps using the architecture and optimization settings described above. Behavioral-model fitting used the recorded human and LLM trajectories and was run as batched cross-validation over agents, reasoning conditions, and folds; because the fitted model has only a small number of parameters per agent, these fits were lightweight compared with LLM generation.

The dominant compute cost came from LLM trajectory generation. Each LLM condition was evaluated on at least $1{,}400$ games, uniformly spanning $T \in \{2, \ldots, 15\}$ sampling rounds, for more than $55{,}000$ LLM games in total. Base-model evaluations constrained the response to a single token, CoT evaluations allowed up to 500 tokens, and extended-reasoning evaluations used the provider's reasoning-effort interface when available. Closed-source models and open-weight models that were too large or costly to host locally were evaluated through the OpenRouter API, so the exact backend hardware for those calls was not under our control.

CPU-only post-processing, statistical tests, and plotting were run on the same workstation/server environment and did not require specialized accelerators. Exact wall-clock times, CPU memory peaks, and total API spend were not retained in the experiment logs; rerun time therefore depends on cluster queueing, API throughput, and provider-side model availability.

E. Task Performance across Rounds

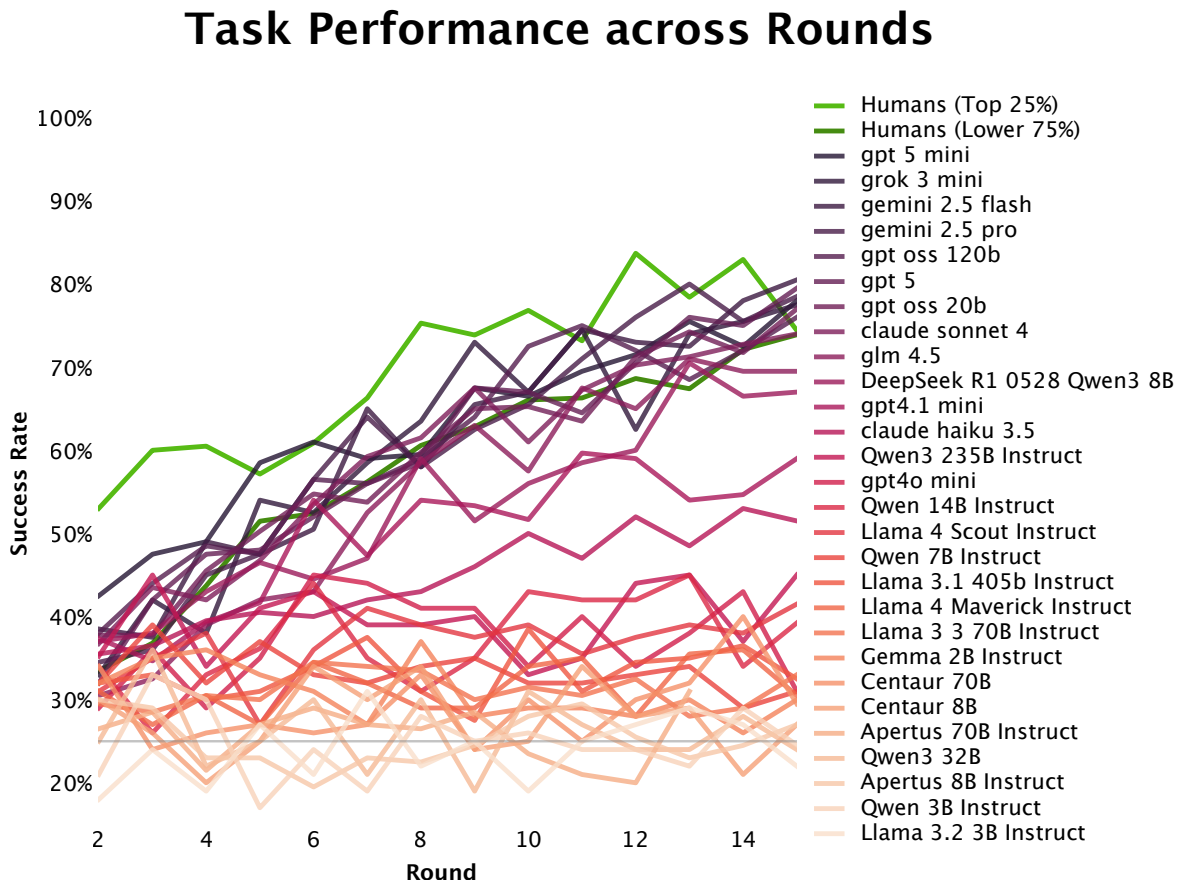

Figure 7. **Evolution of success rate across rounds.** For each agent, we report the mean success rate (fraction of trials in which the final reported button is the true biased one) as a function of the number of sampling rounds $T \in \{2, \ldots, 15\}$. Humans are shown in green, split into the top 25% and lower 75% of participants (same split as in Fig. **2A**); LLMs are shown as individual traces (legend), with colors indicating each model's overall average success rate (lighter/orange = lower, darker/purple = higher). Across models, this view reveals a qualitative separation around $\sim 45\%$ average success: below this regime, curves remain approximately flat with increasing T, suggesting limited ability to convert additional evidence into improved final decisions; above it, curves exhibit a clear positive slope ("lift-off"), indicating effective inference in longer games. *Claude haiku 3.5* is the first model in the performance ranking to show this lift-off, and from *DeepSeek R1 0528 Qwen3 8B* onward, many models track the round-by-round improvement profile of the lower 75% human cohort. Overall, models that display lift-off are predominantly those with extended reasoning.

F. Human Participants

We collected human data in a 1-hour, live in-person competitive session. Under the authors' institutional policy, this low-risk behavioral data collection was not subject to formal IRB review. The study was conducted under institutional ethical standards, and all participants provided written informed consent after receiving task, risk, and compensation information. Fifty participants played independent games of our task with trial length sampled uniformly from $T \in \{2, \ldots, 15\}$ under the same instructions used throughout the paper (Fig. 1A). Forty-six participants completed the full protocol of 100 games each; all analyses in the main text therefore use these $46 \times 100 = 4{,}600$ trials.

Figure 8 reports participant-level performance and decomposes it into inference versus sampling components, complementing the aggregate human summary in Fig. 2. Human success rates span a wide range (roughly 37%–77%; chance $= 25\%$), with most participants clustered around $\sim 55\%$–67%. This heterogeneity is informative for our central comparison: the human distribution overlaps the best LLMs in overall success, while the strongest participants remain a meaningful reference point for what robust inference *and* effective sampling look like in this task.

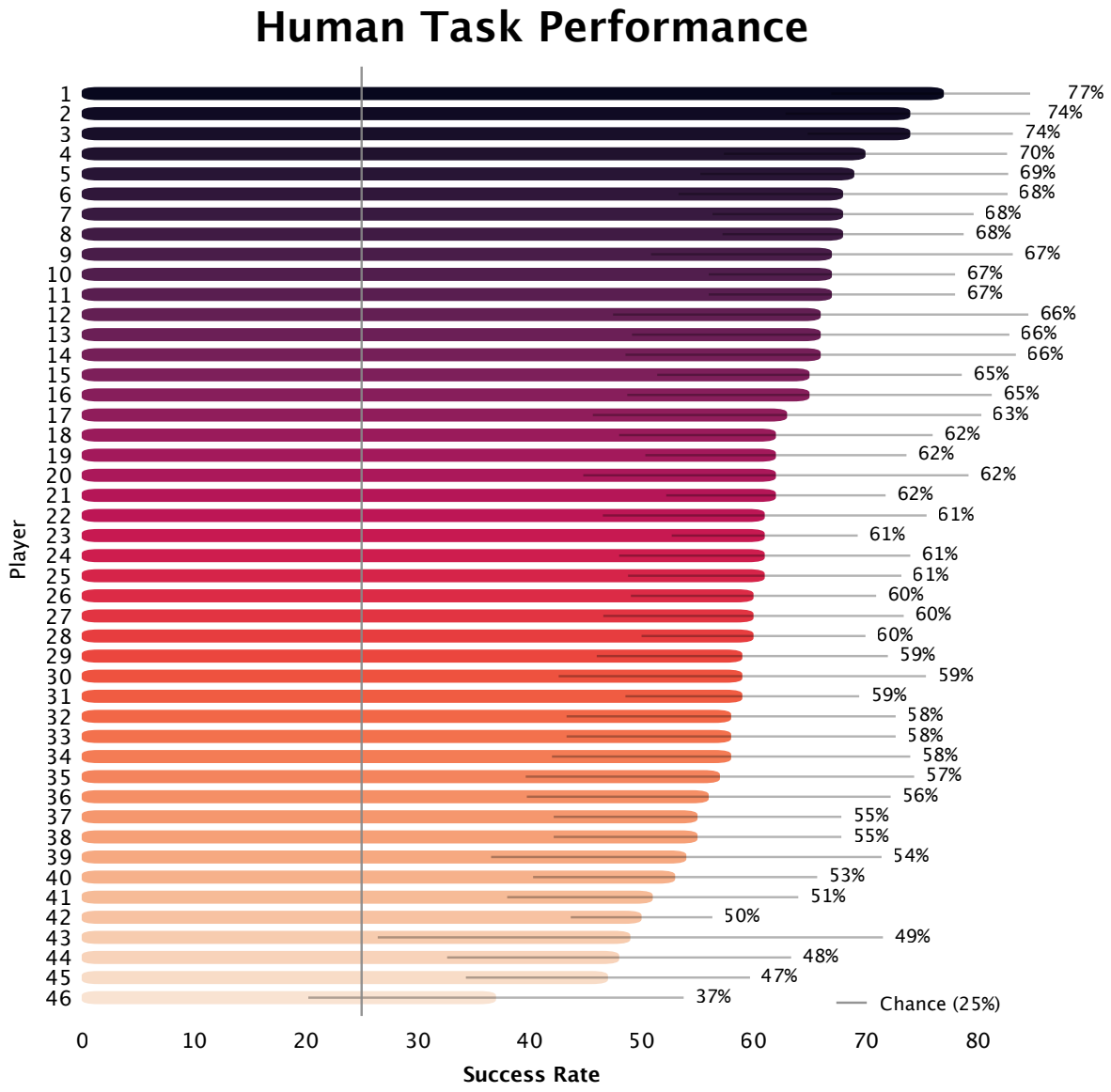

Figure 8. **Individual human performance.** Each row corresponds to one of the 46 participants who completed 100 games (trial lengths uniformly sampled from $T \in \{2, \ldots, 15\}$). Participants are ordered by mean success rate. Mean success rate (correct identification of the biased option at the final decision); the vertical line marks chance performance (25%). The cohort exhibits substantial heterogeneity, spanning $\approx 37\%$ to $\approx 77\%$ success, with most participants concentrated around ~ 55–67%.

G. Behavioral Model reproduces Behavioral Metrics

Model evaluation is performed by simulating the fitted model policy on the *Active Probabilistic Reasoning Task*. For each model and reasoning condition parameters, we roll out full games with the same underlying $\{\alpha_B, \alpha_U\}$ parameters as in the Human and LLM experiment. At each round, the internal memory representation is updated from the initial posterior and a sampling policy is formed from the *choice bias* and *internal posterior*. The comparative posterior/bias plots can be seen in Fig. 9 and represent the average evolution of the underlying posteriors across games with $\{2, ..., 15\}$ rounds, split by sampling and inference components.

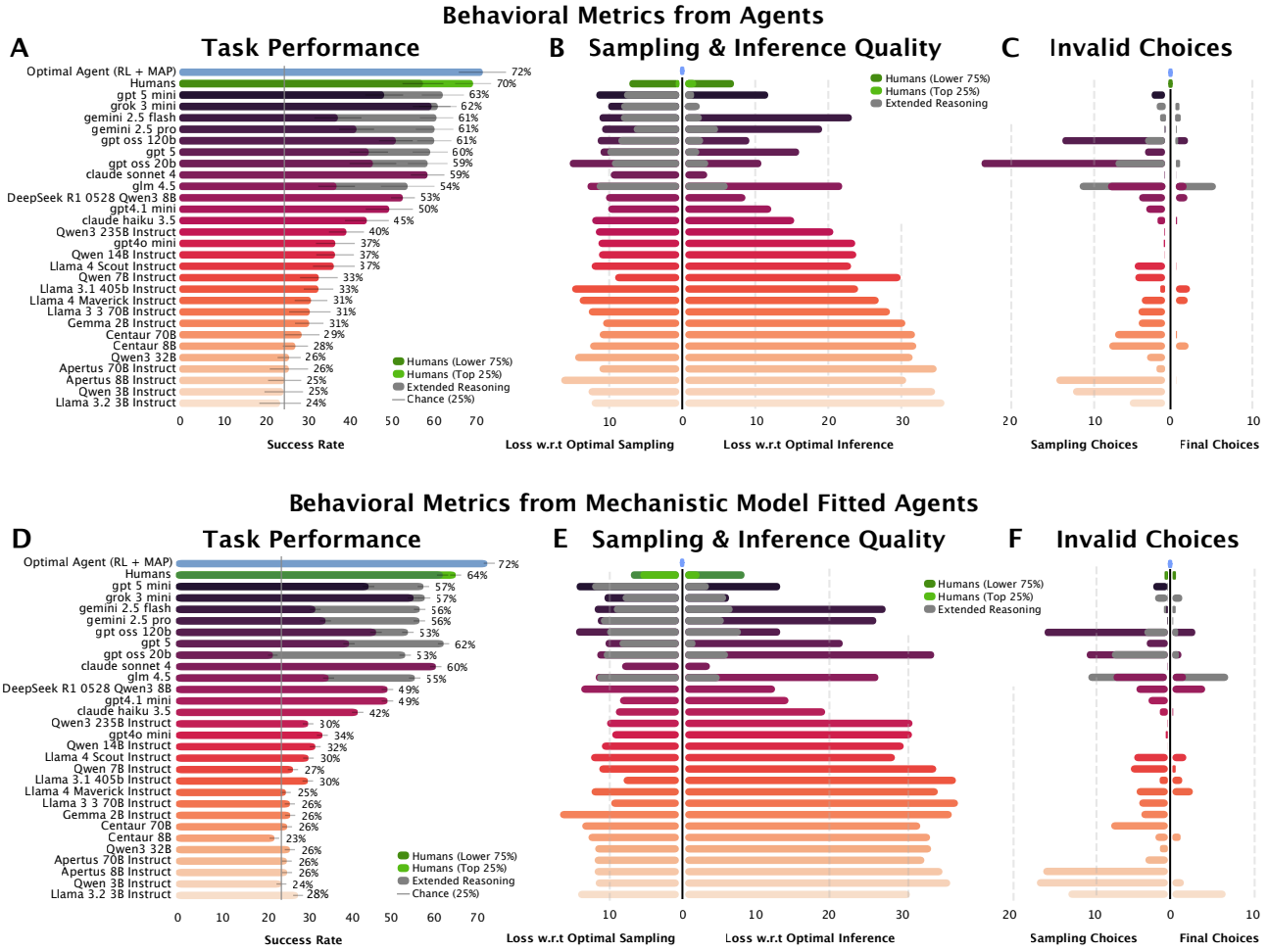

Figure 9. **Behavioral metrics are recapitulated by the fitted behavioral model. A–C**: Empirical behavioral metrics computed directly from human and LLM trajectories. **A**: Overall task performance (success rate) across agents. **B**: Decomposition of performance loss into *sampling* versus *inference* deficits, measured by comparing each agent to counterfactual and optimal reference policies (loss w.r.t. optimal sampling on the left; loss w.r.t. optimal inference on the right). **C**: Invalid-choice rates during sampling and final report. **D–F**: The same metrics computed from trajectories generated by the *behavioral model* after fitting $(\beta, \kappa, \omega, \theta)$ to each agent. **D**: Model-predicted success rates closely match empirical performance ordering. **E**: Model reproduces the sampling–inference loss decomposition, capturing that reasoning primarily reduces inference loss while leaving a residual sampling gap. **F**: Model reproduces invalid-choice profiles via the occlusion-awareness component. Green markers denote human quartiles; grey indicates extended-reasoning conditions.

H. Per model parameter values

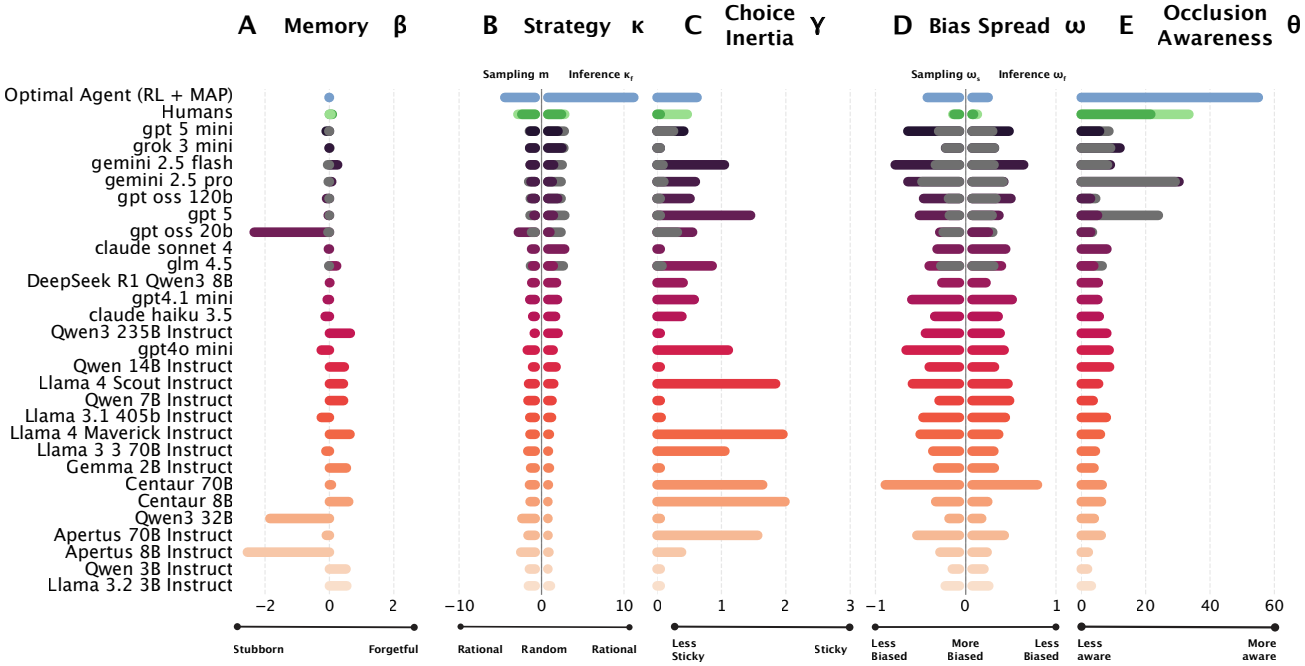

Figure 10. **Model parameters explain human and LLM behavior. A**: Memory β spans *stubborn* ($\beta < 0$) to *forgetful* ($\beta > 0$) evidence integration. **B**: Strategy separates sampling gain m in the sampling temporal kernel $k_s(t)$ from final-choice sharpness κ_f. **C**: Choice inertia γ captures perseveration on the most recent valid sampling choice. **D**: Bias spread (difference between max and min ω_x bias vector components) summarizes systematic sampling and inference choice preferences (ω_s, ω_f). **E**: Occlusion awareness θ captures avoidance of invalid or occluded choices. **F**: Evidence updates posterior p_t and memory h_t; strategy, bias, inertia, and awareness map beliefs to policy π_t.

I. Activation extraction and linear probing

Activation extraction. For each of 50 games per *reasoning level*, we ran *gpt oss 20b* (24 transformer blocks, $d_{\text{model}} = 2880$) through the full task prompt and recorded the residual-stream activations after the MLP of each block at every round (15 sampling rounds plus 1 inference round). Because reasoning traces vary in length and structure across rounds, we summarized each round's activations by averaging over the last 5 tokens of the trace, following standard practice for token-position-robust probing in autoregressive models. The resulting tensor has shape 50 games $\times$ 16 rounds $\times$ 24 layers $\times$ 2 reasoning levels $\times$ 2880 dims.

Linear probes. We fit ridge regression probes that predict the model's choice at each round from its activations, separately for each (layer, reasoning level) pair. Following Alain & Bengio (2016); Belinkov (2022); Gurnee & Tegmark (2023), we use the closed-form ridge solution with the regularization strength α selected by leave-one-out cross-validation over a logarithmic grid $\alpha \in \{10^{-3}, \ldots, 10^{6}\}$. We targeted the choice as a 4-way categorical over $\{A, B, C, D\}$ and report held-out R^2 averaged across 5 K-fold cross-validation folds (uniform-average over the 4 target dimensions).

Swapped control. A high probe R^2 on activations alone does not tell us whether the decoded choice subspace is specific to the reasoning condition or simply reflects task-level structure shared across conditions (e.g., token identity in the trace). To test for condition-specificity, we trained probes on one condition and evaluated them on the other — denoted *X=high, Y=low (SC)* and *X=low, Y=high (SC)* in Fig. **4D**. SC R^2 collapses to near zero or below across all layers, confirming that the choice subspaces are condition-specific.

J. PCA geometry of layer-1 activations

Conditional-average dataset. A naive PCA over all individual-game activations would mix variance from games, layers, and conditions, obscuring the round-by-round structure of interest. To preserve this structure while suppressing game-level noise, we constructed a conditional-average dataset: for each (round, choice, reasoning level, layer) combination, we averaged activations across the games in which the model made that choice on that round. With 16 rounds, 4 choices, 2 conditions, and 24 layers, this yields a $16 \times 4 \times 2 \times 24 = 3072$-row matrix in $\mathbb{R}^{2880}$. We restrict our analyses to layer 1 (selected for its high probe R^2; Appendix **I**) and to a single reasoning level at a time, leaving 128 rows per PCA fit, but the PCA basis is well-defined at any layer.

Variance explained. The top four PCs capture the bulk of the variance in both reasoning conditions (Fig. **11A–B**): under *low reasoning*, PC1 alone accounts for 27.8% of the variance and the top four PCs together explain $\sim 64\%$, while under *high reasoning* the variance is more evenly distributed across components, with PC1 capturing 19.6% and the top four PCs reaching $\sim 61\%$. The flatter spectrum under *high reasoning* is consistent with richer round-by-round dynamics that occupy a higher-dimensional manifold, whereas the steeper spectrum under *low reasoning* reflects trajectories that remain confined to a narrower subspace.

Centroids. For each (choice, reasoning level) pair, we compute a *centroid* as the mean of its conditional-average vectors across rounds, in the original 2880-dimensional space. We use these centroids both as projection targets in PCA visualizations (Fig. **4E–F**) and as references for high-dimensional distance computations.

Centroid-distance similarities. For the example game in Fig. **4**, we compute Euclidean distances d_k from each round's 2880-dimensional activation to all four choice centroids (within the matching *reasoning* condition), and convert them to a similarity distribution via $p_k \propto \exp(-d_k)$. This places the resulting 4-vector on a simplex enabling direct visual comparison between the model's latent variables and a quantity computed entirely from raw activations (Fig. **4A**, bottom row of each panel). For visualization, we additionally project per-round activations into the layer-1 PCA basis and color the trajectory by each round's *closest centroid*, i.e. the $\arg\min$ over the same distances d_k (Fig. **4E–F**).

Pairwise PC views. Three-dimensional projections can occasionally distort relationships between centroids that are well-separated in higher dimensions. As a robustness check, we provide pairwise 2D projections across the top four PCs (Fig. **11C** for *high reasoning* and Fig. **11D** for *low reasoning*); the qualitative pattern: *high-reasoning* trajectories migrating from *C* to *A*, *low-reasoning* trajectories remaining near *C*, is consistent across PC pairs.

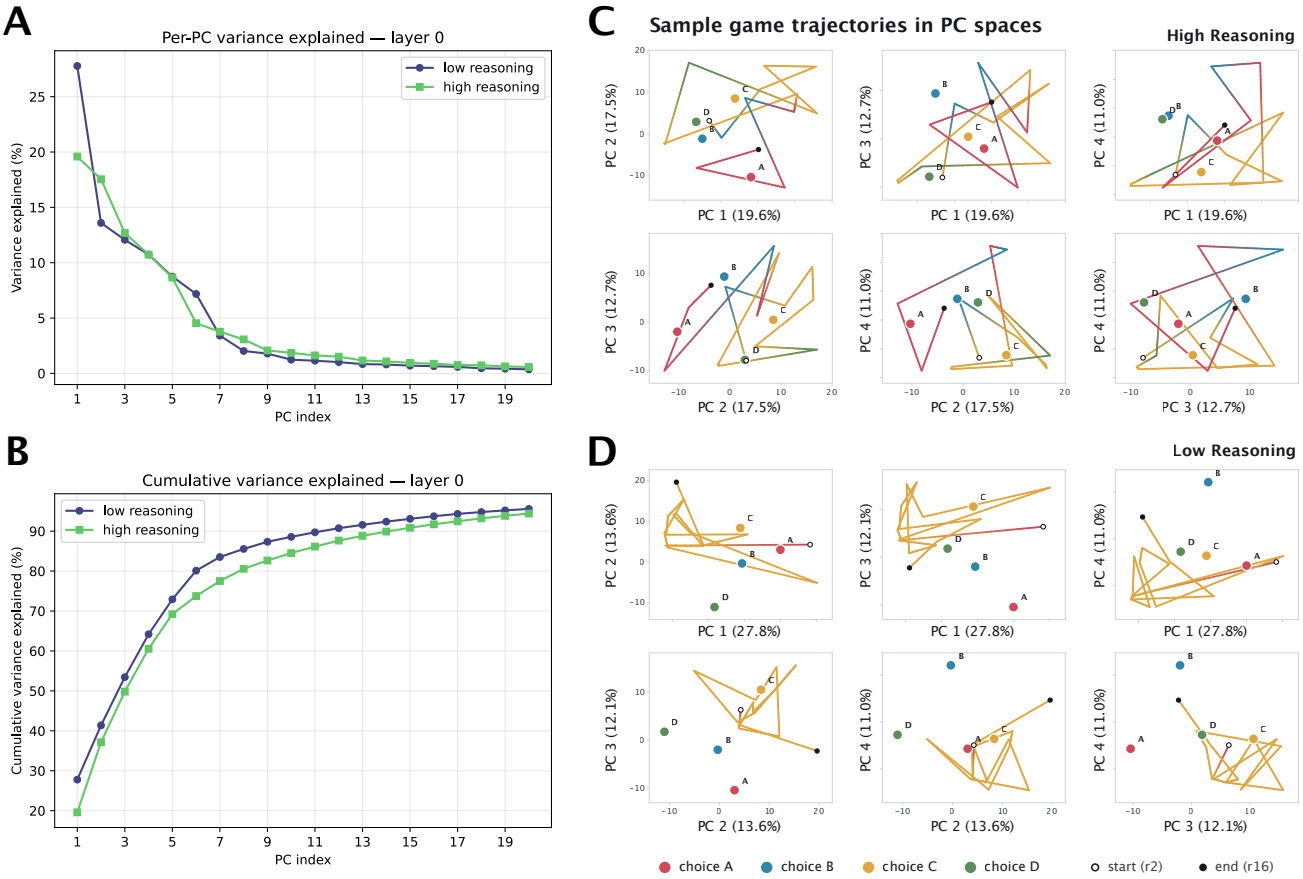

Figure 11. **PCA geometry of layer-1 activations. A:** Per-PC variance explained for the conditional-average dataset at layer 1, shown separately for *low* (purple) and *high* (green) reasoning. PC1 captures 27.8% of the variance under *low reasoning* and 19.6% under *high reasoning*; the flatter spectrum under *high reasoning* indicates richer round-by-round dynamics distributed across more components. **B:** Cumulative variance explained for the same data; the top four PCs together account for $\sim$64% (*low*) and $\sim$61% (*high*) of the total variance. **C–D:** Pairwise 2D projections of the example game's layer-1 activations (same game as in Fig. 4) across the top four PCs, under *high reasoning* (**C**) and *low reasoning* (**D**). Per-choice centroids are shown as filled circles colored by choice (A: red, B: blue, C: yellow, D: green); the trajectory's start (round 2) and end (round 16) are marked by open and filled black circles. Across all PC pairs, the *high-reasoning* trajectory consistently migrates from the neighborhood of centroid *C* toward centroid *A* as evidence accumulates, whereas the *low-reasoning* trajectory remains anchored near *C* throughout the game.

K. Model Formulation Details

We define the behavior model as a structured set of departures from an ideal Bayesian observer. Agents may integrate evidence imperfectly, show systematic choice preferences, ignore occlusions and make invalid choices, or translate beliefs into choices with either MAP-like exploitation or near-random exploration. The goal is to characterize each fitted agent by how evidence is accumulated over time and how the resulting beliefs are converted into sampling and final choices.

K.1. Task likelihood and Bayesian reference

We consider independent games indexed by g. Game g contains T_g[2] sampling rounds followed by one final inference choice. There are K possible cues, with $K = 4$ in the task. Each game has a latent biased cue $z_g \in \{1, \dots, K\}$ with prior $\boldsymbol{p}_0 \in \Delta^K$, $(\boldsymbol{p}_0)_k = \Pr(z_g = k)$. At sampling round t, the agent chooses $a_{g,t} \in \{1, \dots, K\}$ and, when the response yields evidence, observes $x_{g,t} \in \{0, 1\}$, where 1 denotes RED and 0 denotes GREEN.

The ideal observer assumes the shared-rate Bernoulli emission model

$$p(x_{g,t} \mid a_{g,t} = i, z_g = k) = \begin{cases} \mathrm{Ber}(x_{g,t}; \alpha_B), & k = i, \\ \mathrm{Ber}(x_{g,t}; \alpha_U), & k \neq i, \end{cases} \qquad 0 < \alpha_U < \alpha_B < 1. \tag{19}$$

Conditional on z_g and the choices, observations are independent across rounds:

$$p(x_{g,1:t} \mid a_{g,1:t}, z_g = k) = \prod_{u=1}^{t} p(x_{g,u} \mid a_{g,u}, z_g = k). \tag{20}$$

For a single game, suppress g when clear and define the likelihood vector

$$(\boldsymbol{L}_t)_k := p(x_t \mid a_t, z = k). \tag{21}$$

The Bayesian posterior is updated by elementwise multiplication and normalization,

$$\boldsymbol{p}_t = \frac{\boldsymbol{p}_{t-1} \odot \boldsymbol{L}_t}{\mathbf{1}^\top (\boldsymbol{p}_{t-1} \odot \boldsymbol{L}_t)}. \tag{22}$$

Equivalently, with unnormalized log-posterior

$$\boldsymbol{\ell}_t := \log \boldsymbol{p}_0 + \sum_{u=1}^{t} \log \boldsymbol{L}_u, \qquad \boldsymbol{\ell}_t = \boldsymbol{\ell}_{t-1} + \log \boldsymbol{L}_t, \tag{23}$$

and softmax $\sigma(\boldsymbol{v}) := \exp(\boldsymbol{v})/[\mathbf{1}^\top \exp(\boldsymbol{v})]$, the posterior is

$$\boldsymbol{p}_t = \sigma(\boldsymbol{\ell}_t). \tag{24}$$

This follows because $\exp(\boldsymbol{\ell}_t)$ is the vector of Bayes numerators. The softmax also satisfies $\sigma(\boldsymbol{v} + \alpha \mathbf{1}) = \sigma(\boldsymbol{v})$ for any scalar α, so only relative log-beliefs affect the posterior.

K.2. Centered evidence and gamma memory

The shared-rate model implies that, on each round, only the chosen cue receives discriminative evidence relative to the shared mismatch likelihood. Define

$$L_t^+ := p(x_t \mid a_t, z = a_t), \qquad L_t^- := p(x_t \mid a_t, z \neq a_t), \tag{25}$$

and the single-step log-likelihood ratio

$$\delta_t := \log \frac{L_t^+}{L_t^-} = x_t r + (1 - x_t) g, \qquad r := \log \frac{\alpha_B}{\alpha_U}, \quad g := \log \frac{1 - \alpha_B}{1 - \alpha_U}. \tag{26}$$

[2]note that in the main text we drop the game index g for simplicity.

If $e_{a_t} \in \mathbb{R}^K$ is the one-hot vector for the sampled cue, then

$$\log \boldsymbol{L}_t = \log(L_t^-)\mathbf{1} + \delta_t e_{a_t}. \tag{27}$$

Because softmax is invariant to common shifts, we work in centered log-belief space. Let

$$C := I - \frac{1}{K}\mathbf{1}\mathbf{1}^\top, \qquad C\mathbf{1} = \mathbf{0}, \qquad C^2 = C. \tag{28}$$

The Bayesian centered evidence increment is

$$\Delta \boldsymbol{h}_t := C \log \boldsymbol{L}_t = \delta_t \left(e_{a_t} - \frac{1}{K}\mathbf{1} \right). \tag{29}$$

Thus the ideal centered integrator accumulates the same evidence as Bayes, but removes the unidentifiable global baseline. When all increments are accumulated equally, $\sigma(\sum_{u=1}^{t} \Delta \boldsymbol{h}_u) = \boldsymbol{p}_t$ up to the prior-centered initial state.

The fitted model replaces exact accumulation with a temporal memory kernel. This kernel captures order effects in belief updating: older evidence can be discounted, preserved, or overweighted, matching recency and primacy distortions studied in leaky-accumulator and belief-adjustment accounts (Usher & McClelland, 2001; Hogarth & Einhorn, 1992). We use an exponential form because it makes recency and primacy deviations symmetric in log-weight space. For equal $|\beta|$, the log weights have matched slopes but tilt toward the current round when $\beta > 0$ and toward the start of the sequence when $\beta < 0$.

$$\boldsymbol{h}_{g,t} := \sum_{u=1}^{t} w_{t,u}(\beta) \Delta \boldsymbol{h}_{g,u}, \tag{30}$$

with

$$w_{t,u}(\beta) = \begin{cases} \exp[-\beta(t-u)], & \beta > 0, \\ 1, & \beta = 0, \\ \exp[\beta(u-1)], & \beta < 0. \end{cases} \tag{31}$$

Here $\beta = 0$ recovers exact accumulation, $\beta > 0$ produces forgetful recency weighting, and $\beta < 0$ produces stubborn primacy weighting.

K.3. Internal posterior and policy

Following bounded-rationality formulations with information-processing costs (Ortega & Braun, 2013), choices are generated from softmax-sharpened internal posteriors. For the selected gamma fit, sampling sharpness depends on how many valid observations are available before the choice:

$$n_{g,t}^- := 1 + \sum_{u<t} \mathbf{1}\{a_{g,u} \text{ valid}\}, \qquad \kappa_s(g,t) := m \, \frac{n_{g,t}^-}{T_g}. \tag{32}$$

This schedule makes sampling initially exploratory and increasingly belief-guided as usable evidence accumulates, consistent with exploration–exploitation accounts (Daw, 2011; Gershman, 2019).

Let $\boldsymbol{h}_{g,t}^-$ be the pre-choice memory state at sampling round t. The sharpened sampling and final-choice posteriors are

$$\boldsymbol{q}_{g,t}^s := \sigma\big(\kappa_s(g,t)\boldsymbol{h}_{g,t}^-\big), \qquad \boldsymbol{q}_g^f := \sigma(\kappa_f \boldsymbol{h}_{g,T_g}). \tag{33}$$

Here m controls how quickly sampling becomes evidence-guided, while κ_f controls final-choice sharpness.

During sampling, not all cues are available. Let $\boldsymbol{v}_{g,t} \in \{0,1\}^K$ be the availability vector and let $\boldsymbol{i}_{g,t} \in \{0,1\}^K$ be one-hot for the most recent valid sampling choice, with $\boldsymbol{i}_{g,t} = \mathbf{0}$ if no valid choice has occurred. It is clearest to write the cue-response policy in log-score space, where the belief, bias, inertia, and occlusion terms add; $\log \boldsymbol{\omega}_s$ and $\log \boldsymbol{\omega}_f$ are elementwise logs of positive bias vectors:

$$\boldsymbol{\eta}_{g,t}^s := \kappa_s(g,t)\boldsymbol{h}_{g,t}^- + \log \boldsymbol{\omega}_s + \gamma \boldsymbol{i}_{g,t} - \theta_s(\mathbf{1} - \boldsymbol{v}_{g,t}), \qquad \boldsymbol{\pi}_{g,t}^s := \sigma(\boldsymbol{\eta}_{g,t}^s), \tag{34}$$

$$\boldsymbol{\eta}_g^f := \kappa_f \boldsymbol{h}_{g,T_g} + \log \boldsymbol{\omega}_f, \qquad \boldsymbol{\pi}_g^f := \sigma(\boldsymbol{\eta}_g^f). \tag{35}$$

Larger θ_s means stronger avoidance of occluded sampling actions, and larger γ means stronger repeat-choice inertia. Equivalently, for any two cues j, k, the sampling and final log odds are

$$\begin{aligned}\log \frac{(\boldsymbol{\pi}_{g,t}^s)_j}{(\boldsymbol{\pi}_{g,t}^s)_k} &= \left(b_{g,t,j}^s - b_{g,t,k}^s\right) + \log \frac{(\boldsymbol{\omega}_s)_j}{(\boldsymbol{\omega}_s)_k} + \gamma\left(i_{g,t,j} - i_{g,t,k}\right) \\ &\quad - \theta_s\left[(1 - v_{g,t,j}) - (1 - v_{g,t,k})\right], \end{aligned} \tag{36}$$

$$\log \frac{(\boldsymbol{\pi}_g^f)_j}{(\boldsymbol{\pi}_g^f)_k} = \left(b_{g,j}^f - b_{g,k}^f\right) + \log \frac{(\boldsymbol{\omega}_f)_j}{(\boldsymbol{\omega}_f)_k}. \tag{37}$$

Here $\boldsymbol{b}_{g,t}^s := \kappa_s(g,t)\boldsymbol{h}_{g,t}^-$ and $\boldsymbol{b}_g^f := \kappa_f \boldsymbol{h}_{g,T_g}$ are the sampling and final belief log-scores.

The same policy can be written in multiplicative form by exponentiating these log scores. Define

$$\boldsymbol{r}_{g,t} := \exp(\gamma \boldsymbol{i}_{g,t}), \qquad \boldsymbol{o}_{g,t}^s := \exp[-\theta_s(\mathbf{1} - \boldsymbol{v}_{g,t})]. \tag{38}$$

Then Eq. (34)–(35) are equivalently

$$\boldsymbol{u}_{g,t}^s = \exp(\boldsymbol{b}_{g,t}^s) \odot \boldsymbol{\omega}_s \odot \boldsymbol{r}_{g,t} \odot \boldsymbol{o}_{g,t}^s, \qquad \boldsymbol{\pi}_{g,t}^s = \frac{\boldsymbol{u}_{g,t}^s}{\mathbf{1}^\top \boldsymbol{u}_{g,t}^s}, \tag{39}$$

$$\boldsymbol{u}_g^f = \exp(\boldsymbol{b}_g^f) \odot \boldsymbol{\omega}_f, \qquad \boldsymbol{\pi}_g^f = \frac{\boldsymbol{u}_g^f}{\mathbf{1}^\top \boldsymbol{u}_g^f}. \tag{40}$$

The log-score form makes the additive contributions transparent, while the multiplicative form shows the equivalent implementation: exponentiated belief scores are combined with bias, inertia, and occlusion factors before normalization.

K.4. Model loss function

For each agent, the dataset consists of G games. Game g provides sampling choices $\{a_{g,t}\}_{t=1}^{T_g}$, observed outcomes $\{x_{g,t}\}_{t=1}^{T_g}$, sampling occlusion sets $\{\mathcal{A}_{g,t}\}_{t=1}^{T_g}$, and a final inference choice f_g. Given parameters

$$\Theta := (\boldsymbol{\omega}_s, \boldsymbol{\omega}_f, \beta, m, \gamma, \kappa_f, \theta_s, \theta_f), \tag{41}$$

we unroll the memory trajectory deterministically from the observed history, then compute the policies above. In the implementation, θ_s is the fitted sampling occlusion penalty previously named θ_{occ}, and the fitted invalid-response probability is represented through

$$\theta_f = \log \frac{1 - p_{\text{abstain}}}{p_{\text{abstain}}}, \qquad P(\text{invalid/missing}) = \sigma(-\theta_f). \tag{42}$$

For any observed response y and cue policy $\boldsymbol{\pi}$, define

$$P_\Theta(y \mid \boldsymbol{\pi}, \theta_f) = \begin{cases} \sigma(\theta_f)(\boldsymbol{\pi})_y, & y \in \{1, \ldots, K\}, \\ \sigma(-\theta_f), & y \text{ invalid/missing}. \end{cases} \tag{43}$$

The negative cue-response log-likelihood is

$$\mathcal{N}(\Theta) = -\sum_{g=1}^{G} \left[\sum_{t=1}^{T_g} \log P_\Theta(a_{g,t} \mid \boldsymbol{\pi}_{g,t}^s, \theta_f) + \log P_\Theta(f_g \mid \boldsymbol{\pi}_g^f, \theta_f) \right]. \tag{44}$$

The minimized objective is

$$\mathcal{J}(\Theta) = \mathcal{N}(\Theta) + \lambda_\omega \, \mathrm{KL}\big(\boldsymbol{\omega}_s \,\|\, \tfrac{1}{K}\mathbf{1}\big) + \lambda_\gamma \, \gamma^2. \tag{45}$$

Table 4. Selected gamma-fit objective weights and parameter bounds. Alignment losses are ordered as written in the text.

Setting	Symbol	Selected value
Choice likelihood weight	λ_{choice}	1.0
Sampling-bias regularizer	λ_{ω}	0.005
Inertia regularizer	λ_{γ}	0.005
Memory bound	β	$[-3, 3]$
Sampling-gain bound	m	$[0.05, 7]$
Inertia bound	γ	$[0.05, 3]$
Sampling occlusion bound	θ_s	$[0, 60]$
Final sharpness bound	κ_f	$(0, 7)$
Invalid-response logit bound	θ_f	$[0, 60]$

K.5. Fitting Results

For each agent (humans and each LLM) and reasoning mode (*Base*/*Extended*), Appendix Figs. 12–15 visualize (i) posterior trajectories during *inference* and *sampling*, and (ii) choice biases alongside the fitted bias weights $\{\boldsymbol{\omega}_s, \boldsymbol{\omega}_f\}$. The same quantities are summarized as heatmaps in Appendix Fig. 16. Full fitted parameter values are reported in Table K.9 (corresponding to Fig. 3 in the main text).

Posterior curves. All posterior panels compress the full K-dimensional belief vector into a single scalar by reading out the posterior mass assigned to the choice that was actually taken. Concretely, let $\boldsymbol{p}_{g,t} \in \Delta^K$ denote the (normative) Bayesian posterior after t samples in game g. Then for *inference* panels (columns 1–2), the plotted quantity is the posterior probability of the *final chosen option*, evaluated as evidence accumulates across rounds. While for *sampling* panels (columns 3–4), the plotted quantity is the posterior probability of the *sampling action chosen on round* t, evaluated *just before* the sample is taken on that round (so that it reflects the beliefs that informed the sampling decision).

Curves are grouped by trial length $N \in \{2, \ldots, 15\}$. For a fixed N, we compute the scalar trajectory above for every game of length N and then average across games, producing one mean trajectory per N (shown as a family of curves, ordered by N). Thus, each line summarizes the *average posterior dynamics conditional on game length* for *inference* and *sampling*.

True posteriors are computed from the observed trajectories of each agent using the normative Bayesian update (Eq. 1). **Fitted posterior readouts** are generated by unrolling the fitted gamma model and reading out the corresponding sharpened internal posteriors.

Bias panels. The bias panels compare empirical choice frequencies against the fitted bias weights. Empirical bars report how often each option is chosen (separately for sampling and inference), while the fitted bars show the learned bias vectors $\boldsymbol{\omega}_s$ and $\boldsymbol{\omega}_f$ (or their summary statistic, depending on the plot), which capture systematic preferences not explained by the posterior.

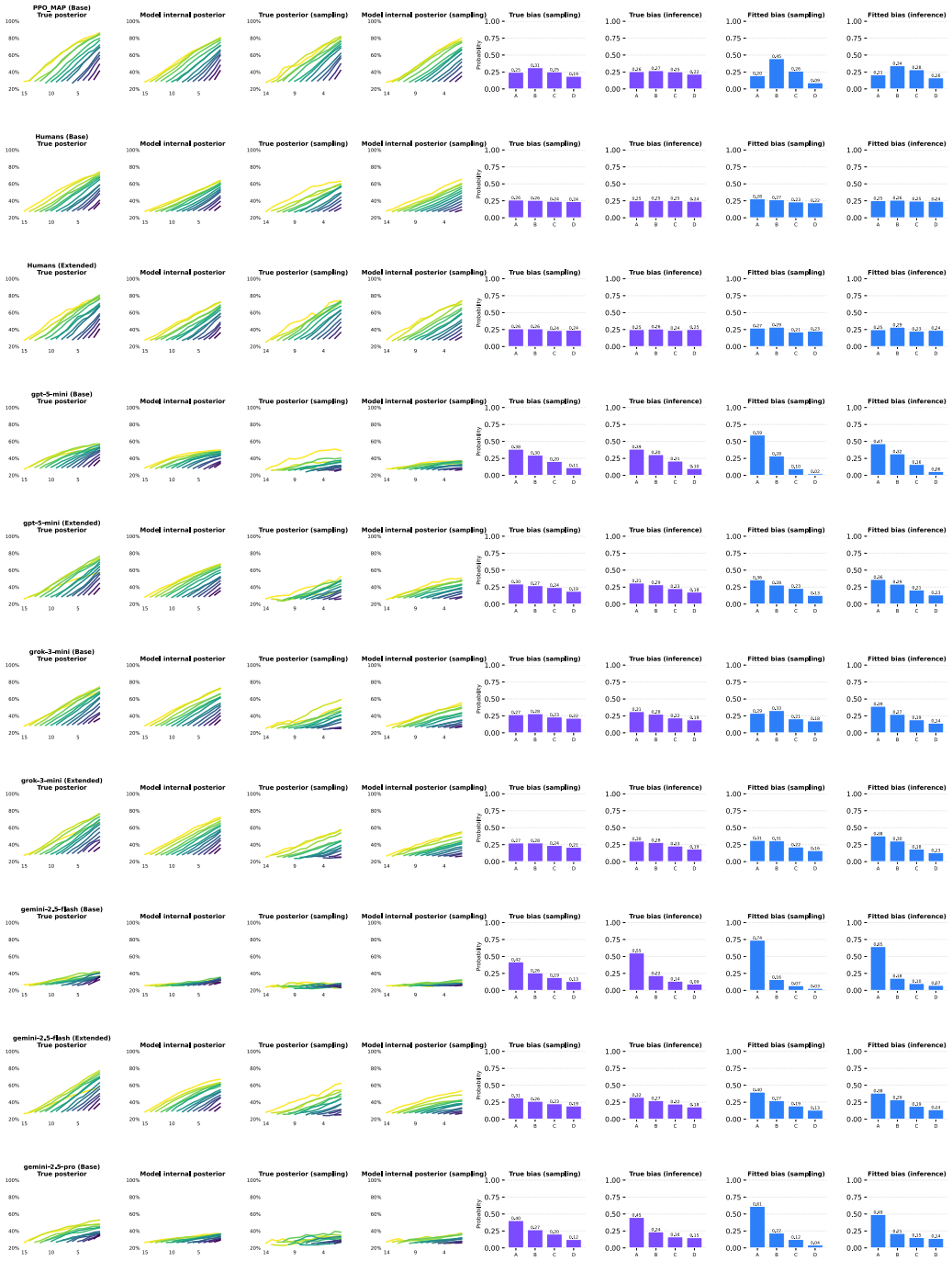

Figure 12. **Posterior evolution by rounds and bias across humans and language models.** In each row the first two panels represent the true and model-generated posterior evolution of the final-choice *inference*, followed by the true and model-generated sampling-choice posterior. The remaining panels show the true bias for sampling and inference and their corresponding fitted bias weights.

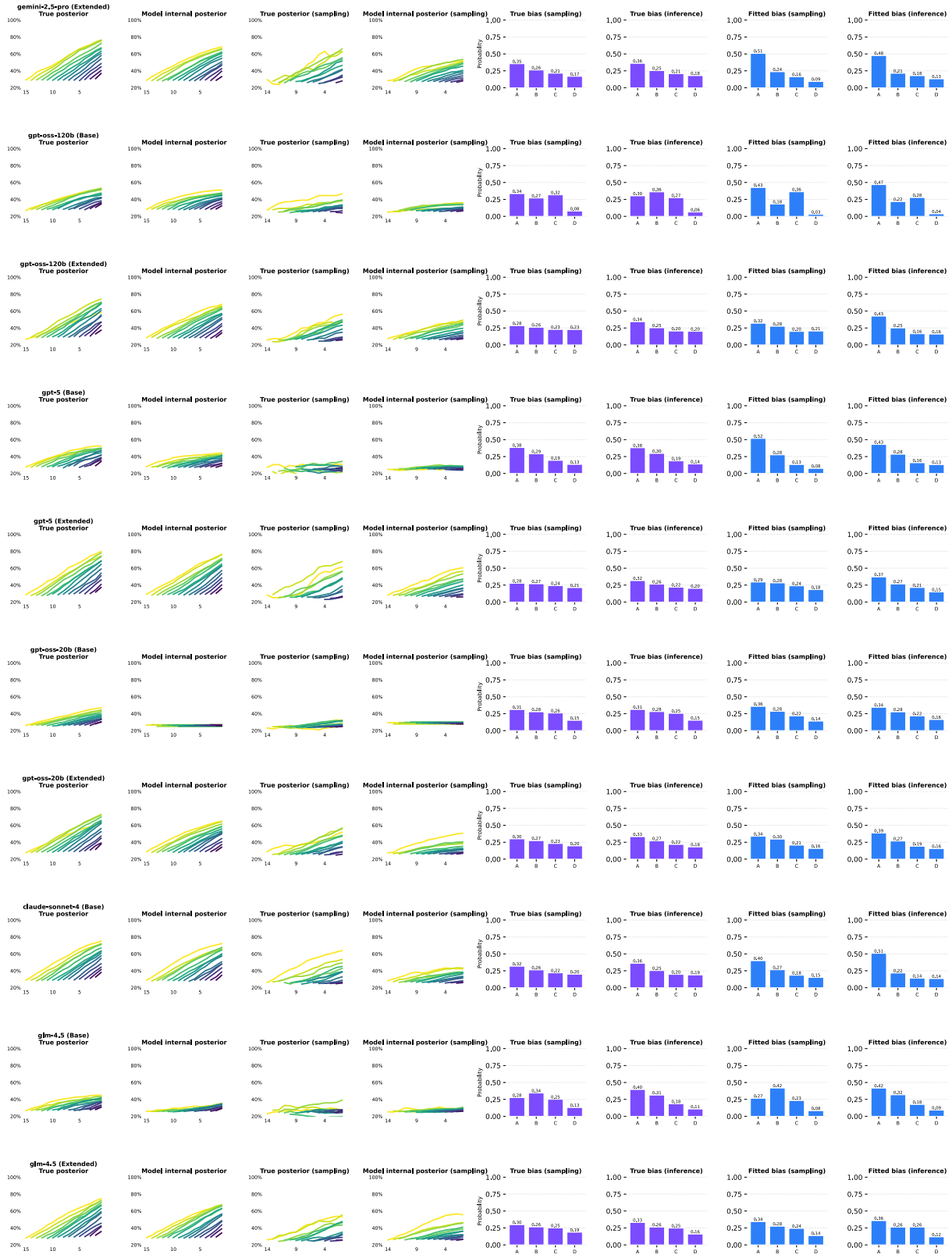

Figure 13. **Posterior evolution by rounds and bias across humans and language models.** In each row the first two panels represent the true and model-generated posterior evolution of the final-choice *inference*, followed by the true and model-generated sampling-choice posterior. The remaining panels show the true bias for sampling and inference and their corresponding fitted bias weights.

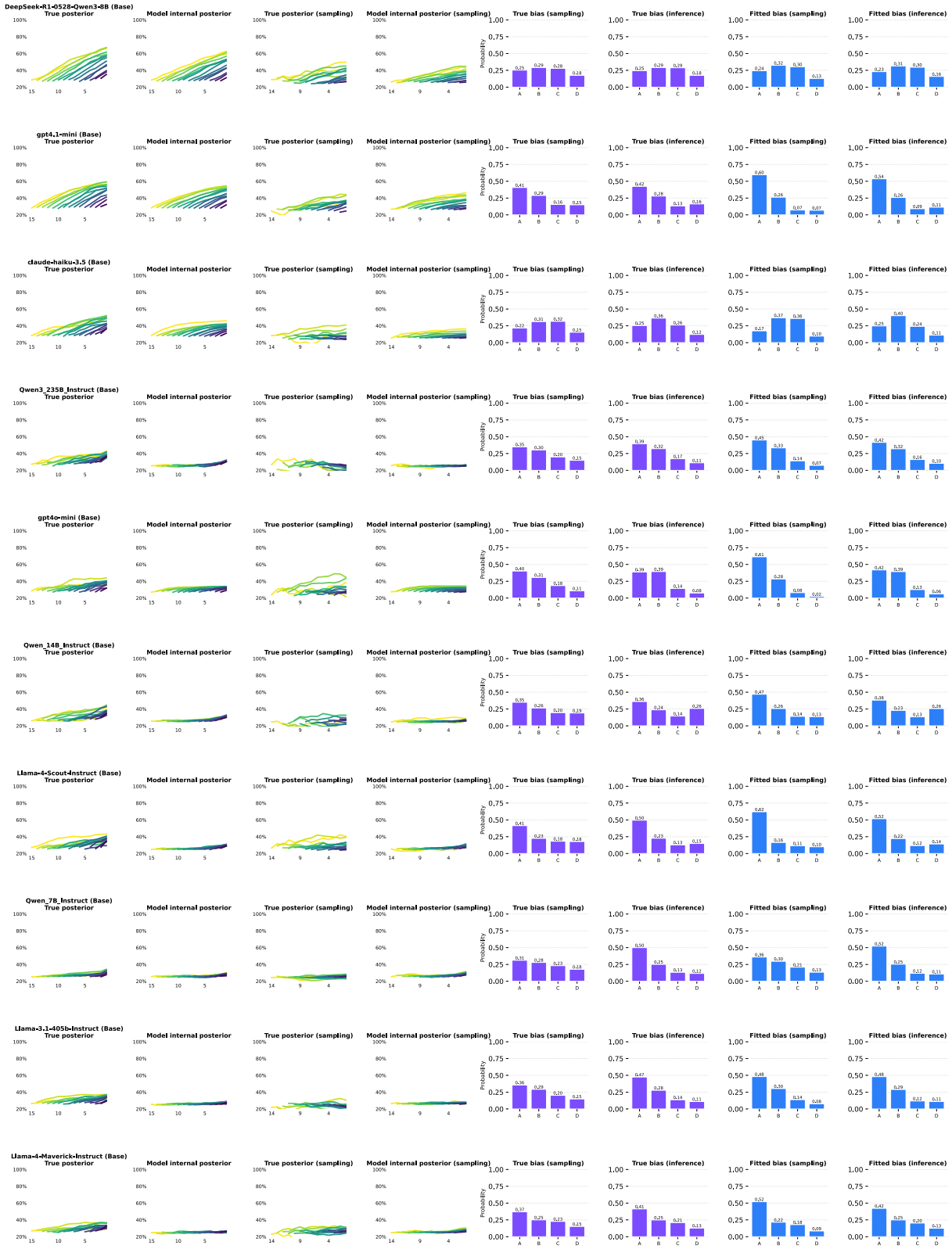

Figure 14. **Posterior evolution by rounds and bias across humans and language models.** In each row the first two panels represent the true and model-generated posterior evolution of the final-choice *inference*, followed by the true and model-generated sampling-choice posterior. The remaining panels show the true bias for sampling and inference and their corresponding fitted bias weights.

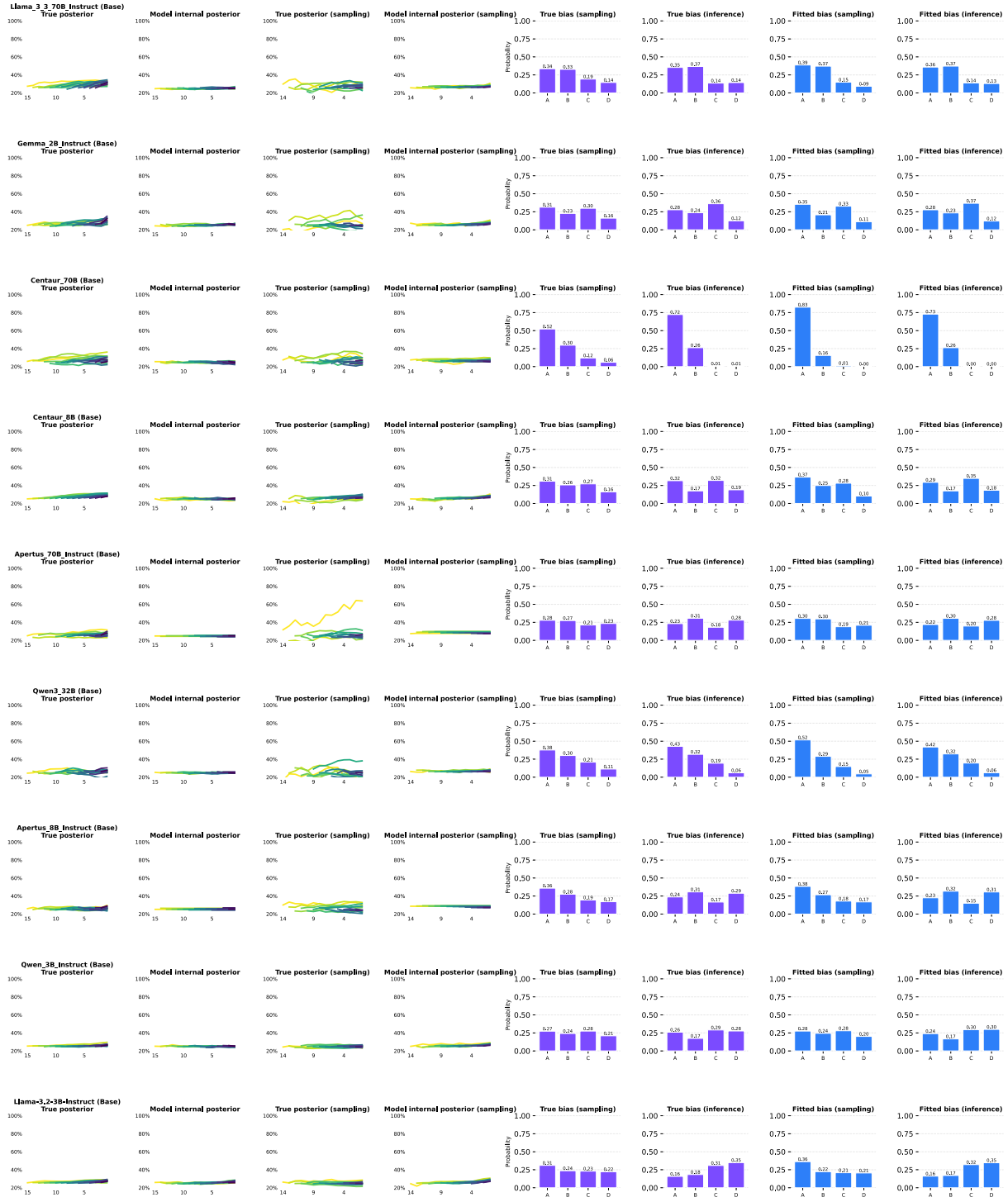

Figure 15. **Posterior evolution by rounds and bias across humans and language models.** In each row the first two panels represent the true and model-generated posterior evolution of the final-choice *inference*, followed by the true and model-generated sampling-choice posterior. The remaining panels show the true bias for sampling and inference and their corresponding fitted bias weights.

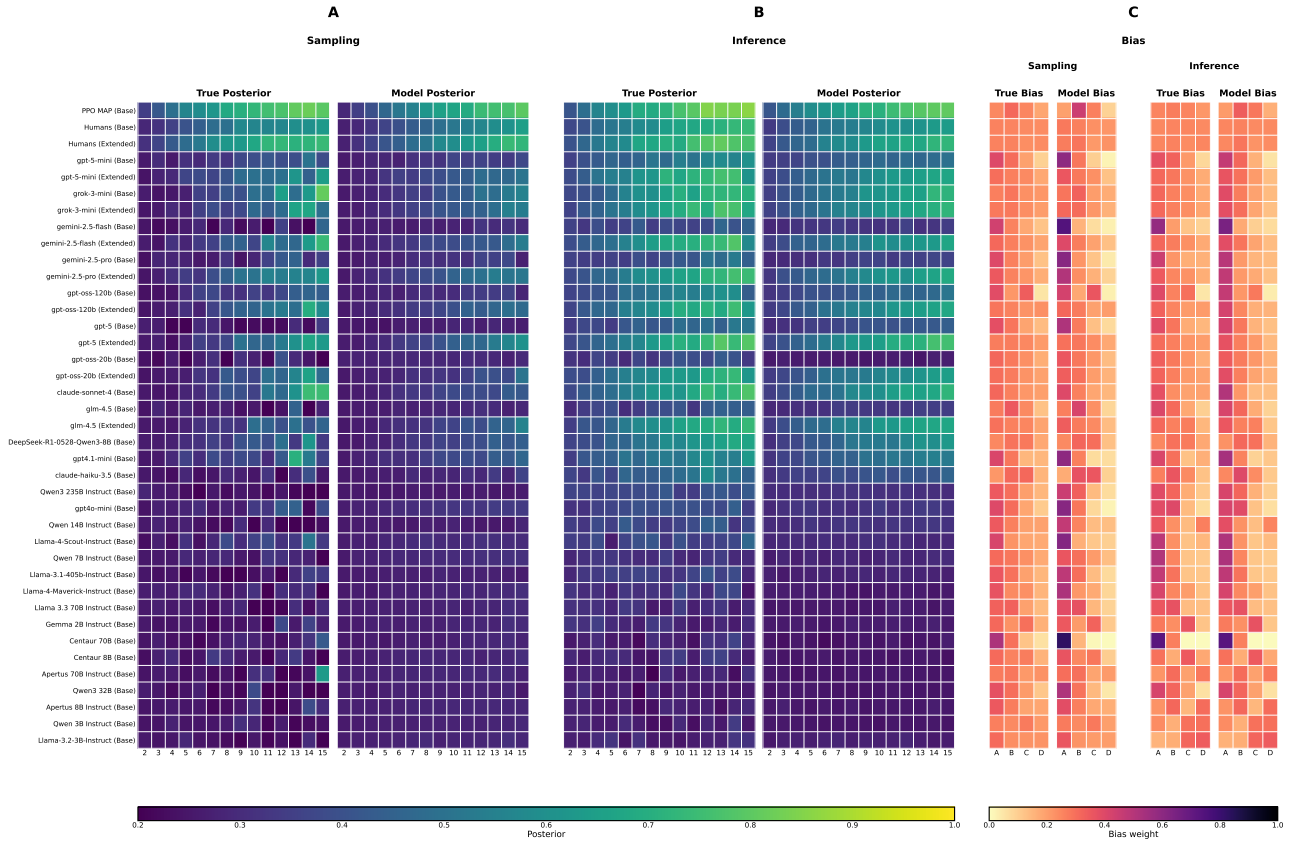

Figure 16. **Posterior and bias structure is captured by the mechanistic model.** For each agent (rows), we compute phase-specific summaries during **sampling** (left block) and **inference** (middle block), and compare empirical estimates (**True**) to quantities generated by the fitted mechanistic model (**Model**). **Sampling**: heatmaps show the agent's **True Posterior** signal (normative posterior mass assigned to the agent's eventual choice under the game history) alongside the corresponding **Model Posterior** obtained by unrolling the fitted belief dynamics. **Inference**: analogous comparison for the final inference phase. **Bias**: right block contrasts **True Bias** (empirical action preference beyond belief) with **Model Bias** recovered from the multiplicative bias parameters. Across agents, the model reproduces both the magnitude and qualitative structure of posterior dynamics and systematic biases, supporting the claim that $(\beta, m, \kappa_f, \omega, \theta_s, \theta_f, \gamma)$ provide a sufficient explanation of the behavioral signatures.

K.6. Statistical significance of behavioral model fits

Table 5. Per held-out game, improvement is $\delta_i = \log p(y_i \mid M) - \log p(y_i \mid B)$, where M is the fitted model and B is an agent-specific frequency baseline. Reported ΔeLPD/game is $\bar{\delta} = \frac{1}{N}\sum_{i=1}^{N} \delta_i$ across out-of-sample games (repeated 5-fold CV; repeated predictions per game averaged first). **CI low/high**: bootstrap confidence interval for $\bar{\delta}$. **p-value**: sign-flip permutation test for $H_0 : \bar{\delta} \leq 0$ vs $H_1 : \bar{\delta} > 0$ (multiple-testing adjusted). **FI**: fraction of held-out games with $\delta_i > 0$.

Model	Reasoning	ΔeLPD/game	CI low	CI high	p-value	FI
PPO_MAP	Base	8.741	8.498	8.997	<0.001	99.1%
Humans	Lower 75%	6.654	6.529	6.779	<0.001	98.7%
Humans	Top 25%	7.275	7.017	7.540	<0.001	97.8%
Base models						
gpt-5-mini	Base	6.962	6.756	7.160	<0.001	98.6%
grok-3-mini	Base	7.354	7.159	7.548	<0.001	98.4%
gemini-2.5-flash	Base	7.716	7.492	7.936	<0.001	99.5%
gemini-2.5-pro	Base	7.239	7.034	7.442	<0.001	99.8%
gpt-oss-120b	Base	3.807	3.664	3.945	<0.001	94.7%
gpt-5	Base	6.065	5.879	6.255	<0.001	98.6%
gpt-oss-20b	Base	3.780	3.580	3.979	<0.001	88.9%
claude-sonnet-4	Base	7.654	7.451	7.854	<0.001	98.9%
glm-4.5	Base	5.911	5.714	6.106	<0.001	97.3%
DeepSeek-R1-0528-Qwen3-8B	Base	6.663	6.457	6.856	<0.001	98.5%
gpt4.1-mini	Base	6.526	6.333	6.728	<0.001	98.3%
claude-haiku-3.5	Base	7.016	6.808	7.223	<0.001	98.8%
Qwen3_235B_Instruct	Base	4.574	4.030	5.090	<0.001	88.8%
gpt4o-mini	Base	7.678	7.449	7.900	<0.001	99.1%
Qwen_14B_Instruct	Base	7.124	6.931	7.318	<0.001	99.2%
Llama-4-Scout-Instruct	Base	6.993	6.770	7.223	<0.001	97.6%
Qwen_7B_Instruct	Base	5.333	5.134	5.527	<0.001	94.2%
Llama-3.1-405b-Instruct	Base	7.288	7.083	7.493	<0.001	99.3%
Llama-4-Maverick-Instruct	Base	6.133	5.931	6.337	<0.001	96.8%
Llama_3_3_70B_Instruct	Base	5.676	5.489	5.871	<0.001	96.5%
Gemma_2B_Instruct	Base	5.402	5.195	5.612	<0.001	95.1%
Centaur_70B	Base	6.149	5.852	6.444	<0.001	92.2%
Centaur_8B	Base	6.249	6.054	6.438	<0.001	98.0%
Apertus_70B_Instruct	Base	5.207	5.030	5.384	<0.001	99.0%
Qwen3_32B	Base	6.569	6.363	6.779	<0.001	97.6%
Apertus_8B_Instruct	Base	2.850	2.690	3.008	<0.001	86.6%
Qwen_3B_Instruct	Base	2.323	2.140	2.505	<0.001	80.1%
Llama-3.2-3B-Instruct	Base	4.860	4.643	5.065	<0.001	92.7%
Extended reasoning models						
gpt-5-mini	Extended	7.805	7.601	8.009	<0.001	99.3%
grok-3-mini	Extended	7.239	7.038	7.439	<0.001	99.0%
gemini-2.5-flash	Extended	7.380	7.189	7.577	<0.001	98.3%
gemini-2.5-pro	Extended	7.694	7.482	7.906	<0.001	98.9%
gpt-oss-120b	Extended	6.589	6.399	6.783	<0.001	97.6%
gpt-5	Extended	7.523	7.325	7.724	<0.001	98.6%
gpt-oss-20b	Extended	5.438	5.259	5.616	<0.001	95.9%
glm-4.5	Extended	6.788	6.594	6.977	<0.001	98.2%

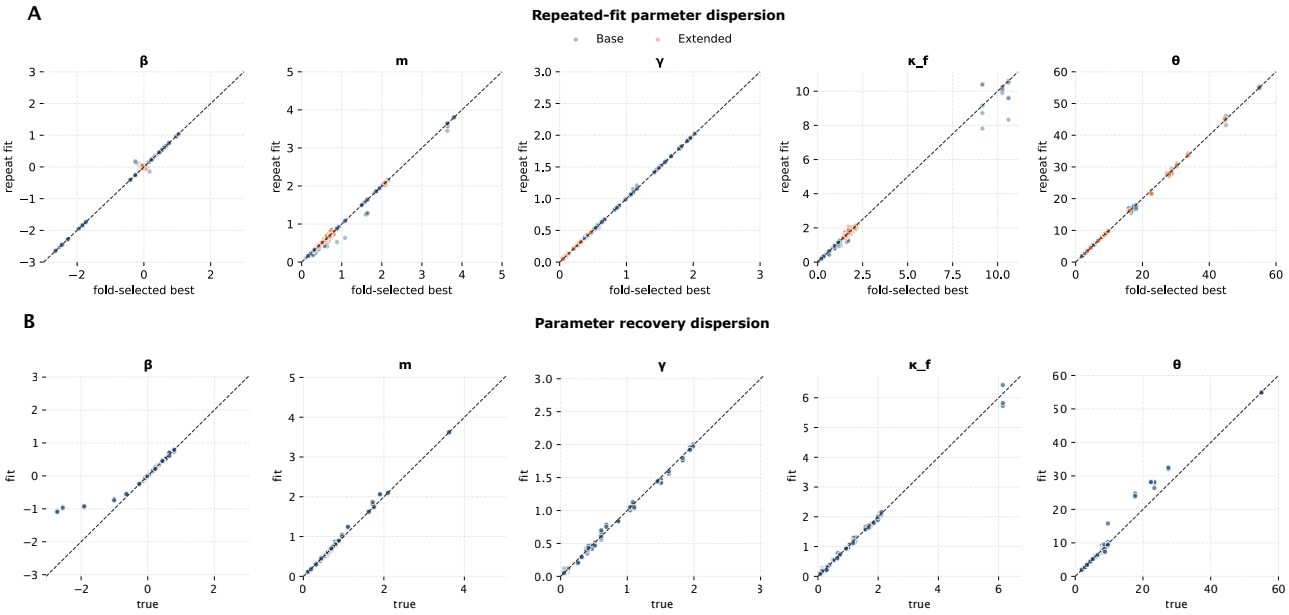

Figure 17. **A**, Fit quality and parameter dispersion across repeated fits. For each agent, the model was fit repeatedly on held-out data splits with multiple random initializations, retaining the best fit within each split by held-out log-likelihood. Each dot is the parameter obtained for one of 10 different parameter initializations. **B**, Parameter recovery for the best model across agents. For each agent, synthetic datasets were generated from the selected fitted parameters. The model was then refit using the same fitting procedure; each dot is a recovered parameter from 10 different initializations.

K.7. Parameter recovery and Parameter dispersion over fits

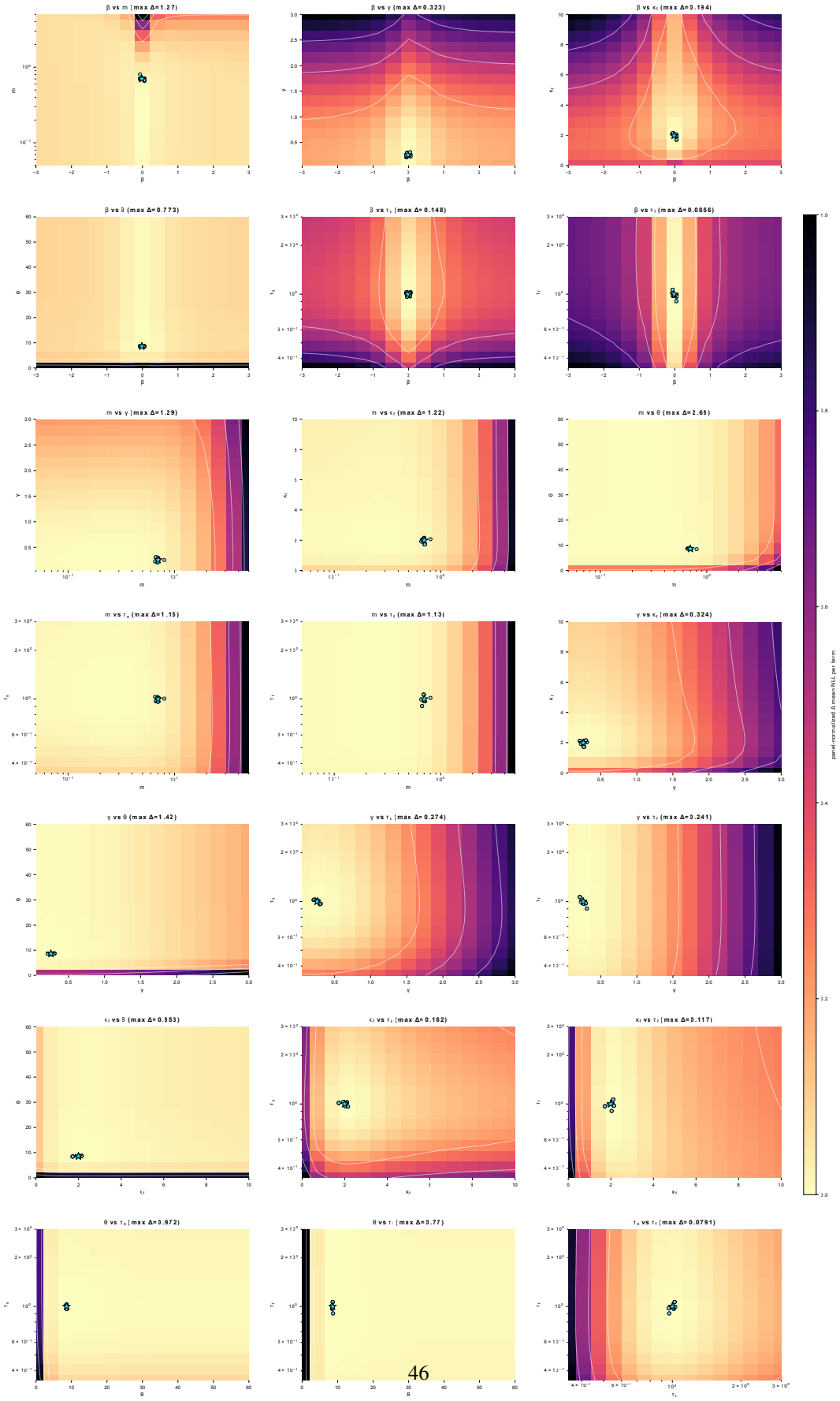

Figure 18. **Pairwise likelihood landscapes *gpt-5-mini Extended*:** All parameters were fixed at their best fitted values except for one pair, which was varied over a two-dimensional grid while evaluating the negative log-likelihood of the empirical data. Each dot shows the resulting parameter values from one of the 10 initializations.

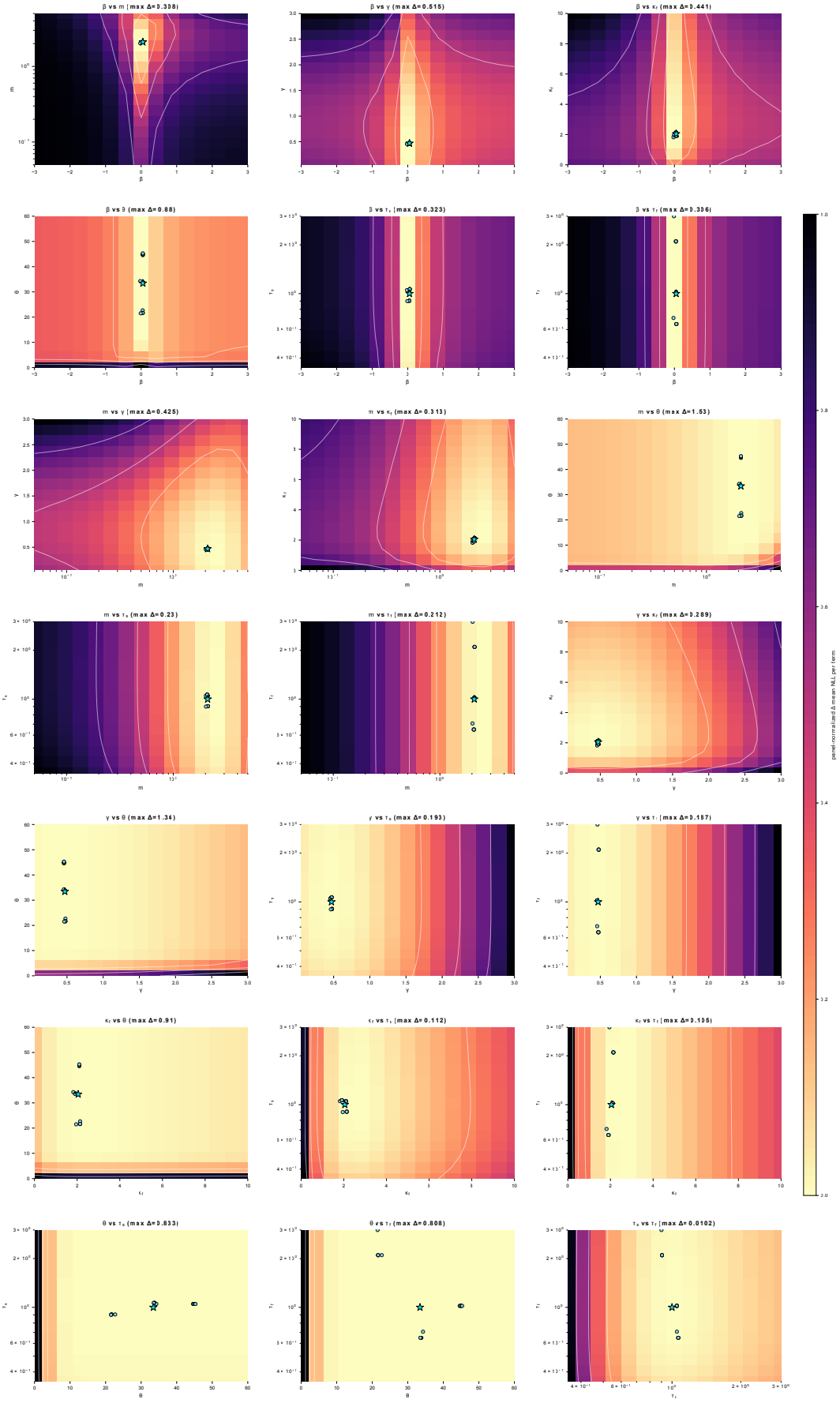

Figure 19. **Pairwise likelihood landscapes *Humans top 25*:** All parameters were fixed at their best fitted values except for one pair, which was varied over a two-dimensional grid while evaluating the negative log-likelihood of the empirical data. Each dot shows the resulting parameter values from one of the 10 initializations.

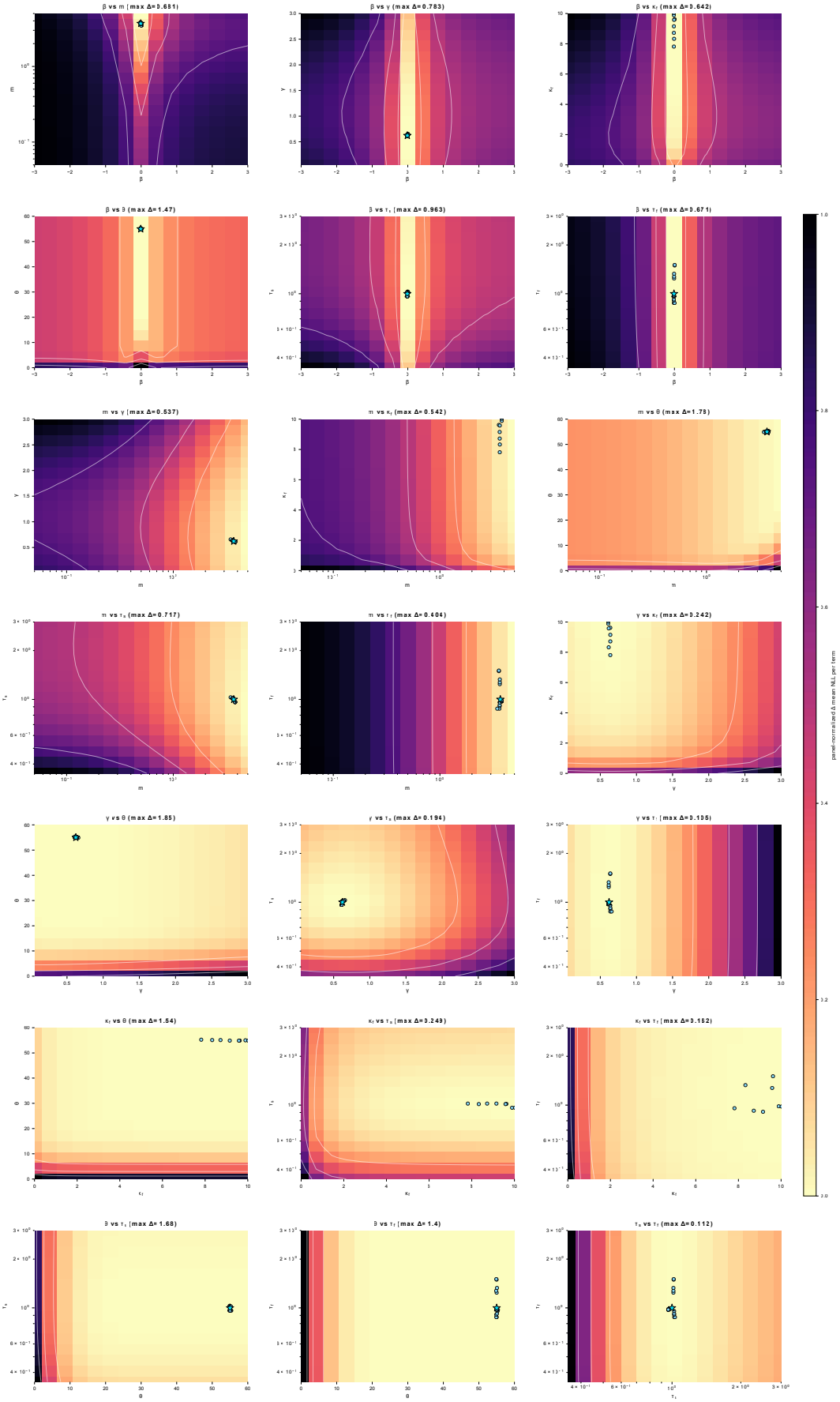

Figure 20. **Pairwise likelihood landscapes *Optimal Agent (PPO)*:** All parameters were fixed at their best fitted values except for one pair, which was varied over a two-dimensional grid while evaluating the negative log-likelihood of the empirical data. Each dot shows the resulting parameter values from one of the 10 initializations.

K.8. Ablation experiments: Structured Parameter Sweeps

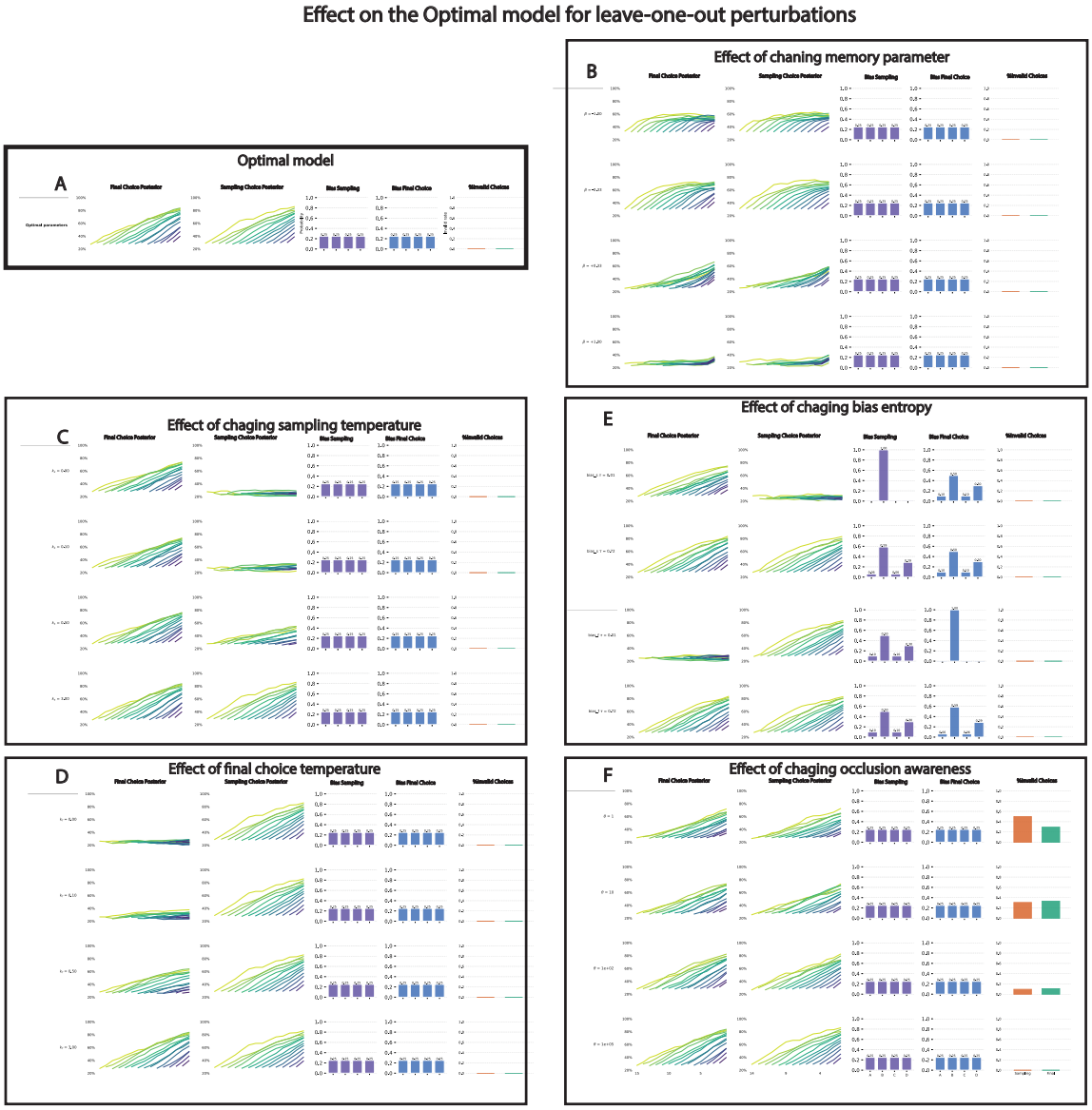

Figure 21. **Parameter sweeps as a proxy for ablation experiments**: Due to the fact that ablating model parameters leads to a particular model assumption, we show here the effect of changing each parameter individually on the observables of the model (Panels B-F) compared to the optimal model parameters (Panel A) with parameters: $\beta = 0$, $m = 3$, $\kappa_f = 3$, $\theta_s = 36$, $P(\varnothing) = 0$, $\mathbf{w}_s = [0.25, 0.25, 0.25, 0.25]$, $\mathbf{w}_f = [0.25, 0.25, 0.25, 0.25]$. **For each row within a panel:** Each row corresponds to one parameter setting, grouped by the parameter being varied (β, m, κ_f, τ_s, τ_f, and θ_s). Here τ is the temperature of the injected bias vector $\mathbf{w}(\tau) \propto \mathbf{w}_0^{1/\tau}$. Within each row, the five panels show: the averaged final-choice posterior summary across rounds, the averaged sampling posterior summary across rounds, the sampling bias weights $\mathbf{w}_s$, the final-choice bias weights $\mathbf{w}_f$, and the invalid-choice rates for sampling and final inference. Posterior summaries are averaged over 100 simulated games for each round setting. Invalid rates are shown only for the occlusion-awareness rows, where θ_f is varied together with θ_s. Changing γ within the $[0.05, 3]$ range has the effect of preventing posterior collapse for high values of m.

K.9. Model Fitting Parameters

Table 6. Fitted parameters of the gamma memory model. Values are best-by-fold averages from the selected fit.

Model	Reasoning	ω_s	ω_f	β	m	γ	κ_f	θ_s
PPO MAP	Base	[0.197, 0.448, 0.264, 0.092]	[0.208, 0.343, 0.284, 0.164]	-0.013	3.687	0.621	10.426	55.000
Humans	Base	[0.280, 0.265, 0.233, 0.221]	[0.253, 0.259, 0.246, 0.243]	0.090	1.626	0.050	1.659	21.587
Humans	Extended	[0.273, 0.286, 0.214, 0.227]	[0.248, 0.286, 0.227, 0.240]	0.048	2.100	0.469	2.048	33.433
gpt-5-mini	Base	[0.594, 0.285, 0.098, 0.023]	[0.467, 0.317, 0.160, 0.056]	-0.101	0.602	0.415	1.230	5.676
gpt-5-mini	Extended	[0.359, 0.281, 0.232, 0.129]	[0.364, 0.294, 0.208, 0.135]	-0.019	0.701	0.257	1.992	8.590
grok-3-mini	Base	[0.289, 0.329, 0.207, 0.176]	[0.392, 0.273, 0.192, 0.143]	-0.012	0.665	0.050	1.709	12.040
grok-3-mini	Extended	[0.313, 0.309, 0.216, 0.163]	[0.379, 0.304, 0.185, 0.132]	0.013	0.675	0.050	1.941	9.119
gemini-2.5-flash	Base	[0.741, 0.160, 0.070, 0.029]	[0.646, 0.179, 0.101, 0.074]	0.255	0.605	1.046	0.699	9.034
gemini-2.5-flash	Extended	[0.399, 0.273, 0.195, 0.133]	[0.384, 0.286, 0.189, 0.140]	-0.056	0.622	0.093	1.740	8.166
gemini-2.5-pro	Base	[0.613, 0.220, 0.125, 0.043]	[0.493, 0.213, 0.154, 0.141]	0.071	0.512	0.595	0.539	30.369
gemini-2.5-pro	Extended	[0.508, 0.237, 0.162, 0.093]	[0.476, 0.214, 0.178, 0.132]	-0.037	0.792	0.106	1.613	29.179
gpt-oss-120b	Base	[0.427, 0.181, 0.363, 0.030]	[0.468, 0.217, 0.278, 0.037]	-0.089	0.420	0.513	1.184	2.879
gpt-oss-120b	Extended	[0.320, 0.275, 0.199, 0.205]	[0.425, 0.251, 0.164, 0.160]	-0.013	0.599	0.050	1.610	4.486
gpt-5	Base	[0.518, 0.275, 0.132, 0.075]	[0.428, 0.284, 0.158, 0.130]	-0.037	0.148	1.452	0.674	5.052
gpt-5	Extended	[0.295, 0.282, 0.239, 0.183]	[0.371, 0.267, 0.214, 0.148]	-0.005	0.625	0.051	2.055	23.948
gpt-oss-20b	Base	[0.358, 0.284, 0.217, 0.141]	[0.344, 0.277, 0.216, 0.163]	-2.337	2.039	0.552	0.248	2.898
gpt-oss-20b	Extended	[0.338, 0.296, 0.207, 0.159]	[0.385, 0.268, 0.190, 0.156]	-0.047	0.446	0.316	1.601	3.475
claude-sonnet-4	Base	[0.399, 0.265, 0.184, 0.152]	[0.508, 0.217, 0.140, 0.135]	-0.032	0.438	0.050	2.038	7.965
glm-4.5	Base	[0.266, 0.418, 0.233, 0.083]	[0.416, 0.318, 0.175, 0.091]	0.224	0.309	0.856	0.660	3.884
glm-4.5	Extended	[0.344, 0.276, 0.244, 0.137]	[0.358, 0.262, 0.262, 0.118]	-0.020	0.569	0.078	1.878	6.503
DeepSeek-R1-0528-Qwen3-8B	Base	[0.244, 0.324, 0.302, 0.130]	[0.230, 0.314, 0.296, 0.159]	0.014	0.418	0.405	1.050	5.307
gpt4.1-mini	Base	[0.596, 0.263, 0.072, 0.069]	[0.536, 0.258, 0.091, 0.114]	-0.068	0.659	0.578	1.170	5.163
claude-haiku-3.5	Base	[0.175, 0.370, 0.359, 0.096]	[0.248, 0.400, 0.242, 0.110]	-0.133	0.305	0.387	0.936	5.599
Qwen3 235B Instruct	Base	[0.451, 0.334, 0.141, 0.074]	[0.416, 0.320, 0.160, 0.104]	0.649	0.113	0.050	1.243	7.934
gpt4o-mini	Base	[0.613, 0.284, 0.080, 0.023]	[0.418, 0.393, 0.126, 0.063]	-0.254	0.961	1.107	0.645	8.667
Qwen 14B Instruct	Base	[0.469, 0.257, 0.140, 0.134]	[0.382, 0.229, 0.133, 0.256]	0.464	0.318	0.050	1.001	8.791
Llama-4-Scout-Instruct	Base	[0.621, 0.165, 0.115, 0.100]	[0.517, 0.223, 0.118, 0.142]	0.435	0.803	1.842	0.714	5.389
Qwen 7B Instruct	Base	[0.360, 0.297, 0.209, 0.134]	[0.522, 0.253, 0.118, 0.106]	0.439	0.875	0.050	0.509	3.713
Llama-3.1-405b-Instruct	Base	[0.481, 0.303, 0.139, 0.077]	[0.480, 0.290, 0.119, 0.111]	-0.252	0.661	0.064	0.446	7.783
Llama-4-Maverick-Instruct	Base	[0.520, 0.215, 0.179, 0.085]	[0.422, 0.249, 0.203, 0.126]	0.640	0.637	1.956	0.186	5.954
Llama 3.3 70B Instruct	Base	[0.387, 0.372, 0.148, 0.092]	[0.359, 0.372, 0.139, 0.131]	-0.109	0.897	1.054	0.069	4.394
Gemma 2B Instruct	Base	[0.354, 0.207, 0.327, 0.112]	[0.277, 0.234, 0.368, 0.121]	0.527	0.557	0.050	0.192	3.966
Centaur 70B	Base	[0.825, 0.156, 0.015, 0.004]	[0.729, 0.263, 0.004, 0.004]	0.062	0.695	1.640	0.066	6.508
Centaur 8B	Base	[0.366, 0.249, 0.281, 0.103]	[0.294, 0.175, 0.349, 0.183]	0.598	0.621	1.987	0.006	6.278
Qwen3 32B	Base	[0.517, 0.289, 0.148, 0.046]	[0.419, 0.323, 0.196, 0.062]	-0.097	0.873	1.564	0.021	6.214
Apertus 70B Instruct	Base	[0.304, 0.298, 0.189, 0.209]	[0.219, 0.305, 0.198, 0.278]	-1.849	1.615	0.050	0.014	4.054
Apertus 8B Instruct	Base	[0.384, 0.268, 0.180, 0.168]	[0.226, 0.318, 0.150, 0.306]	-2.548	1.752	0.381	0.134	2.194
Qwen 3B Instruct	Base	[0.275, 0.244, 0.280, 0.201]	[0.240, 0.166, 0.295, 0.299]	0.516	0.705	0.050	0.005	1.951
Llama-3.2-3B-Instruct	Base	[0.363, 0.224, 0.208, 0.205]	[0.162, 0.170, 0.321, 0.347]	0.539	0.823	0.050	0.314	3.139

Table 7. Negative log-likelihoods of the mechanistic model repeated fold fits. Values are averages over the 15 repeated fold fits per model-condition.

Model	Reasoning	Train NLL	Test NLL	NLL
PPO MAP	Base	0.352	0.353	0.353
Humans	Lower 75%	0.686	0.686	0.686
Humans	Top 25%	0.557	0.559	0.559
gpt-5-mini	Base	0.613	0.615	0.615
gpt-5-mini	Extended	0.653	0.654	0.654
grok-3-mini	Base	0.664	0.666	0.666
grok-3-mini	Extended	0.710	0.712	0.712
gemini-2.5-flash	Base	0.539	0.541	0.541
gemini-2.5-flash	Extended	0.676	0.679	0.679
gemini-2.5-pro	Base	0.628	0.627	0.627
gemini-2.5-pro	Extended	0.610	0.612	0.612
gpt-oss-120b	Base	0.945	0.946	0.946
gpt-oss-120b	Extended	0.796	0.797	0.797
gpt-5	Base	0.712	0.712	0.712
gpt-5	Extended	0.650	0.651	0.651
gpt-oss-20b	Base	1.055	1.059	1.059
gpt-oss-20b	Extended	0.901	0.904	0.904
claude-sonnet-4	Base	0.668	0.669	0.669
glm-4.5	Base	0.912	0.913	0.913
glm-4.5	Extended	0.922	0.924	0.924
DeepSeek-R1-0528-Qwen3-8B	Base	0.846	0.847	0.847
gpt4.1-mini	Base	0.661	0.663	0.663
claude-haiku-3.5	Base	0.750	0.753	0.753
Qwen3 235B Instruct	Base	0.693	0.694	0.694
gpt4o-mini	Base	0.543	0.546	0.546
Qwen 14B Instruct	Base	0.725	0.727	0.727
Llama-4-Scout-Instruct	Base	0.735	0.737	0.737
Qwen 7B Instruct	Base	0.887	0.888	0.888
Llama-3.1-405b-Instruct	Base	0.733	0.734	0.734
Llama-4-Maverick-Instruct	Base	0.750	0.753	0.753
Llama 3.3 70B Instruct	Base	0.799	0.798	0.798
Gemma 2B Instruct	Base	0.875	0.876	0.876
Centaur 70B	Base	0.493	0.491	0.491
Centaur 8B	Base	0.715	0.717	0.717
Qwen3 32B	Base	0.645	0.646	0.646
Apertus 70B Instruct	Base	0.890	0.892	0.892
Apertus 8B Instruct	Base	1.137	1.141	1.141
Qwen 3B Instruct	Base	1.297	1.299	1.299
Llama-3.2-3B-Instruct	Base	1.131	1.132	1.132

K.10. Model-search and selected parametrization

We performed an extensive model search and compared a set of model architectures that kept the same Bayesian evidence variables but changed how memory, sampling sharpness, repeat-choice inertia, and fitting constraints were represented. The goal was not to maximize expressivity, but to find the smallest model that could reproduce the observed behavioral components while leaving the fitted parameters interpretable.

The main alternatives can be written as changes to four components. For memory, a simpler leaky recursion can be defined as

$$\boldsymbol{h}_t = (1-\beta)\boldsymbol{h}_{t-1} + \Delta\boldsymbol{h}_t, \qquad w_{t,u}^{\mathrm{old}}(\beta) = (1-\beta)^{t-u}, \tag{46}$$

whereas the selected model uses the symmetric exponential weights in Eq. (31). For sampling sharpness, we compared the following functionals

$$\kappa_s^{\mathrm{const}}(g,t) = \kappa_s, \qquad \kappa_s^{\mathrm{time}}(g,t) = m\,\frac{t}{T_g}, \qquad \kappa_s^{\mathrm{obs}}(g,t) = m\,\frac{n_{g,t}^{-}}{T_g}. \tag{47}$$

For repeat-choice behavior, no-inertia variants set $\boldsymbol{r}_{g,t} = \mathbf{1}$, while the selected model uses $\boldsymbol{r}_{g,t} = \exp(\gamma \boldsymbol{i}_{g,t})$. Finally,

choice-only fits minimized $\mathcal{N}_{\text{choice}}(\Theta)$ plus simple regularization, whereas the selected fit uses Eq. (45) so that the fitted posterior readouts are also constrained.

Failure modes observed in the search.

Variant family	Mathematical change	Failure mode observed in fits
Simple leaky memory	Eq. (46).	Positive and negative β values were not comparable: recency and primacy did not form mirror-image deviations, making memory attributions unstable.
Constant sampling sharpness	$\kappa_s^{\text{const}}(g,t) = \kappa_s$.	Captured final inference better than sampling dynamics; sampling remained too static within a game.
Round-time sampling gain	$\kappa_s^{\text{time}}(g,t) = mt/T_g$.	Became sharper when rounds elapsed even if the model had produced invalid choices and gained no usable evidence.
No action inertia	$\boldsymbol{r}_{g,t} = \mathbf{1}$.	Repeat-choice streaks were absorbed into β, $\boldsymbol{\omega}_s$, or θ_s, weakening the time dependent interpretation of memory, bias, and occlusion awareness.
Choice-only objective	Omit $\boldsymbol{\lambda}_s^\top \boldsymbol{\mathcal{A}}^s + \boldsymbol{\lambda}_f^\top \boldsymbol{\mathcal{A}}^f$.	Could match cue choices while producing posterior trajectories that missed level, slope, terminal, or rise structure in the Bayesian readouts.

The selected parametrization, Eq. (30)–(45), was the most stable across the full model-search campaign. It reproduced the behavioral components we needed simultaneously: temporal evidence integration through β, sampling dynamics through m, final-choice sharpness through κ_f, cue preferences through $\boldsymbol{\omega}_s, \boldsymbol{\omega}_f$, repeat-choice inertia through γ, and occlusion/nonresponse behavior through θ_s, θ_f.