FROM INTERACTION TO ABSTRACTION: USING BEHAVIOR AND BRAIN ACTIVITY TO EVALUATE HOW AI SYSTEMS LEARN GAMES

Botos Csaba[1]* **Sreejan Kumar**[2,3]* **Austin Tudor David Andrews**[1]

Rui Ponte Costa[1] **Marcelo G. Mattar**[3] **Momchil Tomov**[4]

[1]University of Oxford [2]Columbia University [3]New York University [4]Harvard University

ABSTRACT

Humans rapidly learn abstract knowledge when encountering novel environments and flexibly deploy this knowledge to guide efficient and intelligent action. Can modern AI systems learn and plan in a similar way? Leveraging a unique dataset of human gameplay with concurrent fMRI recordings, we compare model-free and model-based reinforcement learning agents, bayes-optimal model-based agents, and frontier Large Reasoning Model. We evaluate models on task performance and behavioral alignment, then use brain activity as a benchmark to probe the internal representations these models construct. Specifically, using encoding models, we assess how well each system's internal representations predict brain activity in regions previously implicated in theory-based reasoning. We find that the Large Reasoning Model most closely matches human behavioral patterns during game discovery and predicts brain activity in theory-coding regions an order of magnitude better than both model-free and model-based alternatives. Our results shed light on the computational principles underlying human-like rapid learning and planning.

1 INTRODUCTION

Humans are remarkably efficient learners. When faced with unfamiliar circumstances, whether it's a new social situation, an unfamiliar tool, or a brand new videogame, we can rapidly gain enough knowledge of their environments to guide intelligent behavior (Tenenbaum et al., 2011). This knowledge can include, for example, causal information of physical objects, intentional agents, and their interactions (Tsividis et al., 2021; Pouncy & Gershman, 2022). However, learning this knowledge is only half the story; we also use our inferred knowledge to explore efficiently, plan ahead, and act adaptively (Ho et al., 2019; Tsividis et al., 2021). This combination of rapid abstract knowledge acquisition and flexible action guided by this knowledge is a hallmark of human intelligence.

Videogames have become a highly-relevant scientific sandbox for observing and testing this capability (Brändle et al., 2021). Tsividis et al. (2021) designed grid world-style videogames using the Video Game Description Language (VGDL), a framework that specifies games in terms of three core components: objects and their properties (Sprite Set), what happens when objects collide or overlap (Interaction Set), and win/loss conditions (Termination Set). This structure mirrors the compositional, object-oriented format of intuitive theories (Tsividis et al., 2021; Pouncy & Gershman, 2022), making VGDL games an ideal testbed for studying how humans quickly learn about novel environments via interactive experience and act on their new understanding. Tsividis et al. (2021) also introduced the Explore–Model–Plan Agent (EMPA), a theory-based reinforcement learning system that formalizes this style of learning: EMPA performs approximate Bayesian inference over symbolic, program-like descriptions of game dynamics. It infers object types, interaction rules, and win/loss to construct a "theory" of the underlying environment's structure to simulate outcomes and plan actions. EMPA matched human learning efficiency in these games while model-free deep RL

*Equal contribution.

agents required orders of magnitude more experience. Tomov et al. (2023) extended this work by having naive participants learn to play VGDL games while recording their brain activity via fMRI. Using encoding models, they showed that EMPA's inferred theories (the "M" of EMPA) predicted BOLD activity in inferior frontal gyrus (IFG) and other prefrontal regions (e.g. Middle Frontal Gyrus, MFG) significantly better than model-free deep RL, providing evidence that these regions encode abstract representations of the environment like that of EMPA. They additionally identified theory prediction error signals — transient increases in activity when observations violated theoretical predictions — in prefrontal cortex, occipital cortex, and fusiform gyrus (FFG), with effective connectivity patterns consistent with top-down theory predictions and bottom-up error-driven updating.

Videogames have been equally central to the field of artificial intelligence. The era of Deep Reinforcement Learning (Deep RL) began with DQN achieving human-level performance on grid world games (Mnih et al., 2015), and games have remained a guiding benchmark ever since. This line of work has driven the development of model-based Deep RL agents that learn internal representations of environment dynamics to support planning and imagination (Ha & Schmidhuber, 2018; Hafner et al., 2023; Wang et al., 2024). More recently, Large Language Models (LLMs) have become the dominant paradigm in AI (Achiam et al., 2023), demonstrating broad capabilities in language understanding, generation, and even code synthesis. However, these models have faced persistent criticism for their limitations in multi-step reasoning and planning (Valmeekam et al., 2023). In response, researchers have developed Large Reasoning Models (LRMs) that generate explicit reasoning traces, trained with reinforcement learning using verifiable rewards on domains like mathematics and code (Liu et al., 2024).

Can these models learn and plan like humans do? Specifically, can Large Reasoning Models interact with highly dynamic videogame environments to rapidly infer causal structure and flexibly deploy that knowledge to guide action? There have been efforts to evaluate the world modeling capabilities of LLMs and LRMs (Hendriksen et al., 2025; Yang et al., 2024), but such evaluations face a fundamental challenge: abstract knowledge is internal, latent structure that informs further action, which in turn generates new observations that reshape this knowledge. Benchmarks that capture this closed-loop, dual process of learning and acting are scarce. Moreover, behavioral benchmarks alone make it difficult to understand internal computations. Brain activity provides a complementary benchmark that probes the representational structure models construct during learning — what has been termed 'cognitive dark matter,' the computational processes that meaningfully shape behavior yet are hard to infer from behavior alone (Mineault et al., 2026). The VGDL paradigm offers a unique opportunity to address this challenge. Because Tomov et al. (2023) collected both behavioral and neural data from humans learning these games, we can evaluate AI models on two complementary dimensions. If a model performs these games in a human-like way, its internal representations should predict activity in the key regions of interest that encode human abstract knowledge and its behavioral signatures should resemble those of human players. In this paper, we reproduce the results of Tomov et al. (2023) comparing the EMPA model and DDQN, showing that EMPA has more human-like learning trajectories than model-free DDQN agents and that EMPA captures brain activity in humans in key regions better than DDQN. We then evaluate a popular open-source LRM, DeepSeek V3.2 (DeepSeek-AI, 2025), and show that it can capture human-like learning behavior better than both EMPA and DDQN as well as predict brain activity at a performance that is an order of magnitude higher than both EMPA and DDQN, showing evidence that LRMs can indeed learn complex videogames requiring abstract knowledge acquisition like humans.

2 Background

2.1 Dataset

We use the public fMRI + behavior dataset (Tomov et al., 2023) from 32 healthy adult participants (17 male and 15 female; 19–36 years; all right-handed) who learned to play a sequence of grid world games while undergoing fMRI. Each participant completed 6 scanner runs in a single session; each run comprised 3 blocks, and each block consisted of 3 levels of a single game (9 levels per game total). Each level was played for 1 minute total, restarting immediately on the same level whenever an episode ended (win/loss), and runs were organized into three partitions with game/level balance across partitions.

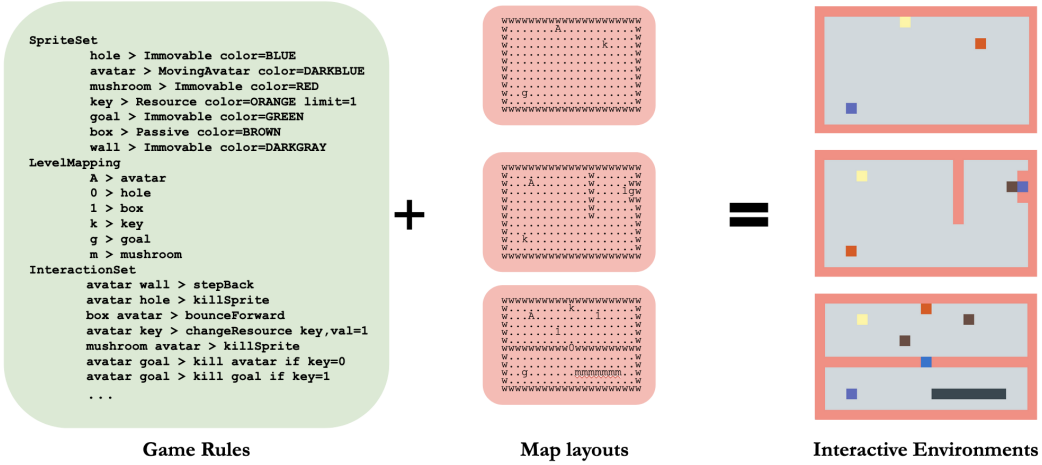

Figure 1: **VGDL Game Structure.** Games are defined by combining game rules (left) with map layouts (center) to produce interactive environments (right). Game rules are expressed in the Video Game Description Language (VGDL, Tsividis et al. (2021)), specifying object types and their properties (SpriteSet), interaction outcomes, and win/loss conditions. Map layouts define the spatial arrangement of objects for each level. Together, these produce the grid-world environments that participants learn to play from scratch. Each participant played 6 games with 9 levels each. Each level was played on repeat for 1 minute, restarting on the same level if the episode ended early. Participants see only the rendered environments with color-only object representations (colored squares) and receive no information about the underlying rules, so they must infer the game's causal structure entirely through interactive experience. Levels are designed so that later levels introduce new object arrangements and interaction opportunities not available earlier, sustaining learning throughout the session.

The games were expressed in Video Game Description Language (Tsividis et al. (2021), Figure 1), which all share a common 5-action gridworld interface (left/up/down/right/action). Levels were explicitly designed to sustain learning: later levels vary layouts and can introduce interaction opportunities not present earlier. To reduce semantic priors, gameplay was presented in "color-only" mode (objects rendered as colored squares with symbols), and participants were instructed that colors/symbols/game names carry no rule information beyond within-game object identity; participants received a base payment plus a performance-dependent bonus tied to the maximum winning score on a randomly selected level.

2.2 Model Classes

Model-Free Deep RL We use Double DQN (DDQN) Van Hasselt et al. (2016) as a model-free reinforcement learning baseline, following the implementation details, architecture, and optimization choices described in Cross et al. (2021); Tomov et al. (2023). In particular, we retain their observation preprocessing, network structure, replay-based training, and exploration strategy, ensuring that our baseline matches the original work as closely as possible in all aspects unrelated to training schedule and model selection.

We modify the training protocol of Tomov et al. (2023) in two key ways. First, instead of training for a fixed number of epochs while cycling through levels in a shuffled order, we train DDQN separately on each level of each game for a fixed budget of 100,000 gradient updates. This allows us to better approximate the dynamic evolution of representations in human participants, capturing the *before* and *after* states of learning new rules on a given level. Second, we explicitly acknowledge the strong sensitivity of deep RL methods to environment-specific hyperparameters. Rather than using a single configuration across all games, we treat each game as an independent tuning problem. For every game, we sample 512 hyperparameter configurations and select the one that maximizes expected performance averaged across that game's levels.

For encoding analyses (see section "Neuroimaging"), we evaluated all internal representations of DDQN: the first two convolutional layers (conv1 and conv2) and the fully connected layer preceeding the penultimate q-values (fc1).

Model-Based Deep RL We additionally evaluate EfficientZeroV2 (EZV2, Wang et al. 2024, a model-based deep RL agent that learns a latent dynamics model to support planning via Monte Carlo tree search. Unlike DDQN, EfficientZero learns an internal transition model of the environment, making it a model-based deep RL counterpart positioned between the purely model-free DDQN and the explicitly symbolic EMPA.

Theory-Based The Explore–Model–Plan Agent (EMPA) Tsividis et al. (2021) model is a non-parametric, model-based baseline that explicitly represents hypothesized causal rules governing object interactions. The agent maintains a structured theory composed of three components: object types and their properties, interaction rules governing what happens when objects collide, and termination conditions specifying win/loss goals. These components are inferred online from observed state transitions via approximate Bayesian inference and are stored in a symbolic memory shared across levels within a game. Planning proceeds by simulating candidate action sequences using the current rule set and selecting actions that maximize expected progress toward terminal reward under these inferred dynamics.

Large Reasoning Model. Large Reasoning Models (LRM) that act via explicit reasoning traces, offer a qualitatively different learning paradigm that is well aligned with the goals of this study. Rather than acquiring task-specific competence through gradient-based training and weight updates (like the RL methods), or explicit access to the rules of the game (like EMPA), these models rely on in-context learning, the ability to infer latent structure, rules, and strategies directly from ongoing interaction and recent experience Brown et al. (2020); Olsson et al. (2022).

This shifts the evaluation setting from one in which models are extensively optimized to overfit a single environment, toward a more human-like regime in which the agent is constrained to the same interaction budget as human participants and must learn purely through transient representations and working memory. As a result, game-specific knowledge in reasoning models is localized in the representational state constructed online (e.g., hypotheses, commitments, and action-conditioned summaries), rather than being diffused across model parameters as in reinforcement learning agents. This makes reasoning models a particularly compelling comparison point for studying representational alignment: if human abstract knowledge is likewise formed and updated online during play, then models that learn exclusively through in-context reasoning may better capture both the dynamics and localization of human neural representations during rapid learning.

3 In-Context Learning Protocol

We evaluate an 685B parameter, open-weight, frontier reasoning model, DeepSeek-AI (2025) as an online hypothesis-testing and planning agent. The agent interacts with the environment step-by-step and receives only the current symbolic state description plus a short summary of recent experience; it performs *no gradient-based training*, parameter updates, or offline dataset fitting. This design isolates the contribution of structured reasoning and tool-like memory from representation learning.

State abstraction and anonymization. At each timestep, the environment returns a structured VGDL state. We transform this state into a text-based, symbolic observation for the LLM using a deterministic formatter. Critically, we hide privileged semantics by anonymizing sprite identities: internal engine identifiers (e.g., `goal.1`, `key.1`) are mapped to abstract IDs `obj_1`, `obj_2`, ... via a one-to-one mapping. The only object attributes revealed to the LLM are those required for interaction: object color, abstract ID, and game-relevant state features (e.g., grid position; plus avatar inventory). In the design of the available information we give exactly the same information that is available for the human participants at any given state of the game.

Single-turn prompting with compact memory. Decisions are produced using a *single-turn* chat call at every timestep. The prompt is constructed from: (i) the current formatted state, (ii) the current level number and attempt counter, (iii) a truncated action history of the last k steps, and (iv) two short

persistent fields carried across steps: `note_to_self` (working hypotheses / discovered rules) and `commitment` (an explicit near-term plan). The LLM is instructed to return a strictly structured object containing, `note_to_self`, `commitment`, and `action`. Importantly, we do *not* provide the full conversational transcript; the agent's only "memory" is the bounded action-history summary plus the two carried fields, which makes context length predictable and comparable across models.

Environment loop For each level, the agent repeatedly: (1) builds the prompt from the current formatted state and recent history, (2) samples a structured response, (3) executes the parsed action in the unwrapped environment, and (4) logs the resulting transition. Levels terminate on win, death, or a fixed step cap (following the same step budget as used in Tomov et al. (2023)). On death, the level is automatically restarted; on win, the agent advances to the next level in the curriculum.

Replay Protocol for Encoding Analysis For the neuroimaging analyses, we use a separate harness designed to replay human trajectories rather than generate agent gameplay. At each timestep, the model receives the same static prompt, the current game state the human participant observed, the history of previous observations, and the action the human actually took. The context is reset at every timestep — the model receives no carry-over memory from previous calls, unlike the generative harness's persistent scratchpad. The model outputs inferred rules and a hypothesized action, but these outputs are discarded for the encoding analysis. Instead, we extract internal representations at the last input token, prior to any model output. Because the input includes the full history of observations alongside the current state, this embedding is a measure of the model's world model, an accumulated (over trial history) understanding of the game's dynamics and structure — rather than the model's downstream reasoning and planning. Because every model receives identical text input for a given human trajectory, differences in brain predictivity across models reflect differences in representational structure.

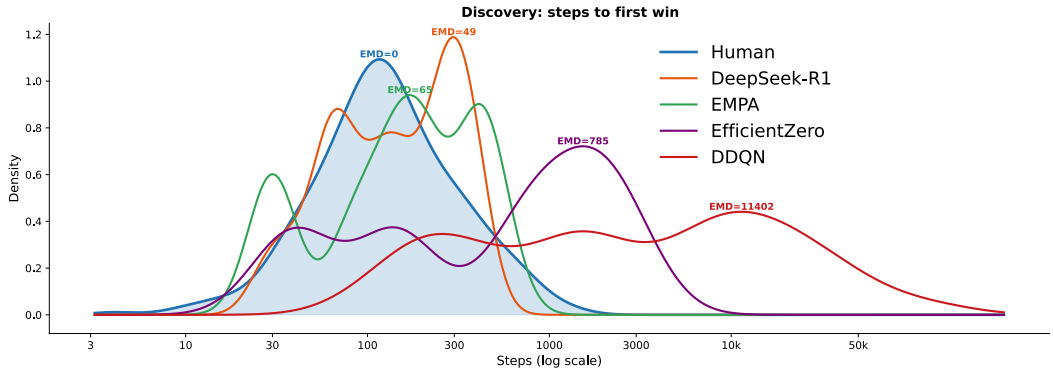

Figure 2: **Discovery behavior: steps to first win across model classes.** Kernel density estimates of steps to first win (discovery phase) for humans and four AI systems across 5 common games, plotted on a log scale. Earth Mover's Distance (EMD) to the human distribution is shown for each model (lower = more human-like). DeepSeek-R1 most closely matches the human discovery distribution (EMD = 49), followed by EMPA (EMD = 65). EfficientZero (EMD = 785) and DDQN (EMD = 11402) require substantially more experience before achieving initial success, consistent with qualitatively different learning regimes. The tight overlap between human and DeepSeek-R1 distributions suggests that in-context learning, under human-like constraints of no task-specific training and no access to game rules, captures the rapid hypothesis-testing characteristic of human game discovery.

4 Results

4.1 Behavior

For each agent class, we compute the distribution of steps to first win on each level, capturing the full exploration cost of discovering a winning strategy. We compare these distributions across humans, EMPA, EfficientZero, DDQN, and the reasoning model to assess whether AI agents require similar amounts of experience as humans to infer game structure and achieve initial success (Figure 2).

Under the most human-like evaluation constraints with no access to game rules, no task-specific training, and no extensive rehearsal, the reasoning model makes the most appropriate comparator to human participants. DeepSeek-R1 and humans exhibit closely overlapping distributions over steps-to-first-win (EMD = 49), consistent with rapid task inference and early identification of goal-relevant structure rather than prolonged trial-and-error learning. EMPA also achieves relatively fast discovery (EMD = 65), though this is harder to interpret as human-like because EMPA benefits from privileged access to the space of possible game rules. EfficientZero (EMD = 785) and DDQN (EMD = 11402) require substantially more experience before achieving initial success, with heavy-tailed distributions spanning orders of magnitude more steps, suggesting qualitatively different learning regimes. The clear separation between model classes, with the reasoning model and EMPA in the human range and the deep RL agents far outside it, suggests that rapid game discovery depends on generic prior knowledge that encapsulates how gaming environments work, whether encoded in a symbolic hypothesis space (EMPA) or acquired through large-scale language pretraining (DeepSeek-R1), rather than on task-specific prior knowledge acquired through gradient-based optimization within the task.

4.2 NEUROIMAGING

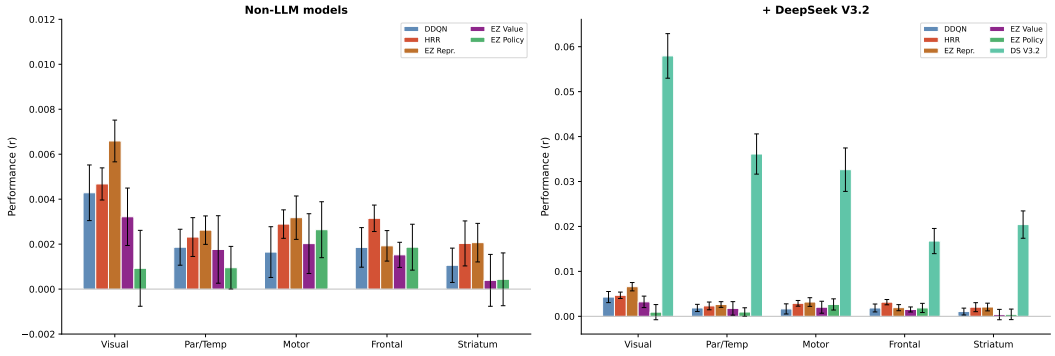

Figure 3: **Encoding model performance by brain region.** (Left) Non-LLM models: DDQN, HRR (EMPA theory embeddings), and EfficientZero split by network component (Representation, Value, Policy). HRR outperforms DDQN across most regions, particularly motor and frontal cortex. EZ Representation layers achieve the highest non-LLM performance in visual cortex, while EZ Value and Policy heads show reduced or negligible encoding. (Right) Same models with DeepSeek V3.2 added. The LLM outperforms all non-LLM models by an order of magnitude across every region group. Best layer selected per region group per model. Error bars indicate ±1 CM-corrected SEM across N=21 subjects (Cousineau et al., 2005).

To assess whether AI model representations align with human brain activity during game learning, we fit encoding models that predict blood-oxygen-level-dependent (BOLD) responses from model features extracted during gameplay.

Encoding Model. Prior to encoding analysis, we parcellated the cortical surface into 1000 regions using the Schaefer atlas (Schaefer et al., 2018), averaging BOLD signal across voxels within each parcel to improve signal-to-noise ratio. We used banded ridge regression to predict parcel-wise BOLD responses from model features (Nunez-Elizalde et al., 2019). This approach assigns separate regularization penalties to different feature groups, allowing the model to appropriately weight heterogeneous predictors. We defined four feature bands: (1) main model features, (2) button press indicators, (3) game and level identity, and (4) temporal variables (time within play, time within experiment). Bands 2–4 serve as nuisance regressors, isolating the unique contribution of model representations. Note that although the original work of Tomov et al. (2023) only controlled for game identity, we take a more strict approach here by also accounting for motor confounds (button presses), autocorrelation (time features), and different level effects (level identity).

To account for the hemodynamic response delay, we created lagged versions of all features at 2, 3, 4, and 5 TRs into the past, concatenating these to form the final design matrix. Features were

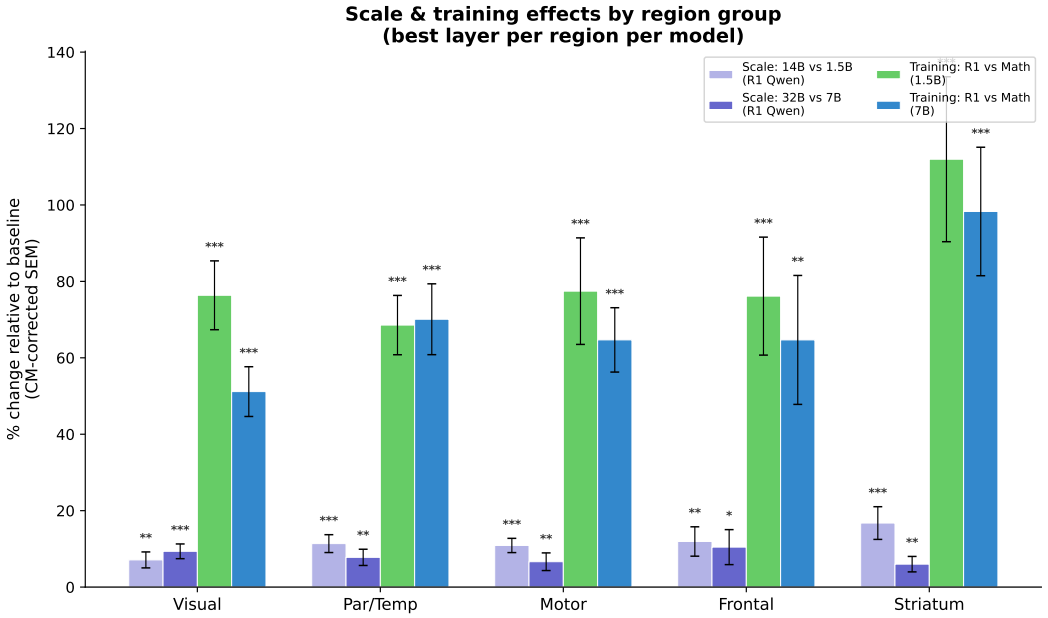

Figure 4: **Reasoning distillation dominates scale in driving brain alignment.** Percent change in encoding performance between matched LLM pairs, averaged within region groups. Purple bars: scale effects (larger vs. smaller R1-distilled Qwen 2.5 models). Green/blue bars: reasoning distillation effects (R1-distilled vs. base pre-distillation model trained on Math). Distillation yields 50–120% improvement across all regions, compared to 4–16% for scale. Striatum shows the largest distillation effect. All models share the Qwen 2.5 base architecture. Each model evaluated at its best layer per ROI. Errorbars indicate SEM across subjects.

z-scored within each cross-validation fold using training set statistics. We applied PCA within each fold to reduce dimensionality. All models were reduced to 230 features, which retained more than 90% variance for every single model we used.

Cross-validation Protocol Following Tomov et al. (2023), we used leave-one-partition-out cross-validation, where partitions correspond to experiment stages: levels 0–2, 3–5, and 6–8. Because later levels introduce more complex rulesets and novel interaction opportunities, this scheme tests generalization across varying degrees of game complexity while ensuring temporal separation between train and test sets. Regularization hyperparameters were selected via nested cross-validation within the training set.

Results We quantified encoding performance as the Pearson correlation between predicted and observed BOLD responses in held-out partitions, averaged across folds. We report results for anatomically-defined regions of interest (ROIs) used in Tomov et al. (2023), spanning frontal/motor, dorsal/parietal, ventral/temporal, and early visual cortex (Figure 3).

We first replicated the qualitative pattern from Tomov et al. (2023): HRR embeddings of EMPA's theory representations outperform DDQN convolutional features, particularly in motor and frontal cortex. We then evaluated EfficientZero, a model-based RL agent, separating its representation network (which encodes spatial game state), value head, and policy head. The representation network achieved the strongest non-LLM performance in visual cortex Adding an LLM, DeepSeek V3.2, produced an order-of-magnitude improvement across all regions of interest, establishing that LLM representations capture substantially more variance in neural activity during gameplay than either EMPA theory representations encoded with HRR or model-free (DDQN) agents, or model-based RL with learned world models (EfficientZero).

4.3 POST-TRAINING DOMINATES SCALE IN DRIVING BRAIN ALIGNMENT

Having established that LLM representations substantially outperform non-LLM models, we next asked what properties of LLMs drive their neural alignment. We isolated two factors: model scale (number of parameters) and reasoning distillation. All models share the same Qwen 2.5 base architecture, but differ in fine-tuning: the R1 variants were distilled from DeepSeek R1's reasoning traces, while the Math variants received supervised fine-tuning for mathematics without reasoning distillation. We computed the percent change in encoding performance relative to the weaker model for each matched comparison (Figure 4). Reasoning distillation effects dwarfed scale effects across every region. Replacing Math with R1-distilled variants at matched size improved encoding by 51–123%, whereas increasing model size within the R1-distilled family (1.5B to 14B, or 7B to 32B) produced only 4–16% gains. This asymmetry was consistent across all five region groups, with the largest distillation advantage in striatum (112% at 1.5B, 123% at 7B). These results suggest that the reasoning capabilities acquired through distillation, not the capacity afforded by additional parameters, is a much more efficient driver of brain-aligned representations during gameplay.

5 DISCUSSION

We evaluated four classes of AI systems — model-free deep RL (DDQN), model-based deep RL (EfficientZero), a Bayes-optimal theory-based agent (EMPA), and a frontier Large Reasoning Model (DeepSeek V3.2) — on their ability to capture human behavior and brain activity during rapid game learning. Behavioral and neural evaluations provide complementary and dissociable insights into these models, demonstrating that brain activity reveals computational properties that behavioral benchmarks alone would miss.

Behaviorally, the LRM most closely matches human discovery play under the most human-like constraints: no access to game rules, no task-specific training, and learning purely within trial. EMPA also falls within the human range, though its privileged access to the VGDL rule space complicates direct comparison. Both deep RL agents require orders of magnitude more experience, consistent with qualitatively different learning regimes. The separation between model classes suggests that rapid game discovery may depend on generic prior knowledge about how gaming environments work, whether encoded in a symbolic hypothesis space or acquired through large-scale pretraining, rather than task-specific prior knowledge learned through gradient-based optimization within the task.

The neuroimaging results reveals information internal to the model that drives behavior Mineault et al. (2026). The LRM's internal world model representations predict brain activity across all regions of interest an order of magnitude better than DDQN, EfficientZero, and EMPA theory embeddings. This cannot be attributed to differences in information content: the replay harness provides every model with identical text input describing the same human trajectory, so differences in brain predictivity reflect how models internally organize that information, not what information they receive. That reasoning-distilled models produce far greater gains in brain alignment than comparable increases in scale (50–120% vs 4–16%) further demonstrates the value of neural benchmarks. The heightened sensitivity of subcortical regions, particularly striatum, to reasoning distillation suggests that what reasoning training instills is specifically relevant to the neural circuits supporting sequential action selection and structured decision-making.

An important caveat applies to the comparison with EMPA. In this study, following Tomov et al. (2023), we compare the modeling component of EMPA: its inferred theory of game dynamics embedded via holographic reduced representations (HRR). EMPA's full architecture encompasses exploration and planning components not tested here. Moreover, EMPA's inferior neural performance may reflect limitations of the HRR embedding scheme rather than its computational content. The symbolic theory EMPA infers could be correct while the format in which it is represented for brain comparison is suboptimal. Embedding EMPA's theories using LRM representations would isolate the contribution of representational format from computational content and is a critical direction for future work.

Several additional limitations should be noted. First, our replay harness captures a measure of the representations of the current world model: an accumulated understanding of game dynamics given the observation history. However, it does not isolate representations associated with exploration,

planning, or action selection. Future work using scaffolds with explicit functional decompositions could map out these different computations. Additionally, we lack a noise ceiling estimate for the fMRI data, making it difficult to determine what fraction of explainable variance each model captures. However, this is a significant methodological challenge given that, unlike perception-only paradigms (e.g. movie watching or image processing), a single play of a game is a unique experience to a participant that cannot be repeated, neither in the same nor different participant. Future work should also complement encoding models with representational similarity analyses to assess whether these results hold across methodological approaches, and should include visual model baselines (e.g., SOTA convolutional networks processing game frames) to further isolate the contribution of language-based representations from visual features.

More broadly, our results highlight the value of using brain activity as a benchmark for understanding what computational properties drive human-like learning in AI systems. The most striking example is the dissociation between scale and reasoning distillation. Increasing model parameters produces modest gains in neural alignment (4–16%), while reasoning distillation at matched scale produces 50–120% improvement, with the largest effects in striatum. That the brain distinguishes between these manipulations, and does so most strongly in subcortical circuits involved in structured sequential behavior, suggests that reasoning training instills representational structure specifically relevant to the neural computations supporting rapid learning and action selection. As games become increasingly central to evaluating general intelligence in AI, through benchmarks like ARC-AGI-3 and AI GameStore (Ying et al., 2026), neural benchmarks offer a complementary evaluation axis that can reveal what models learn internally, beyond what their outputs alone can show.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Franziska Brändle, Kelsey R Allen, Josh Tenenbaum, and Eric Schulz. Using games to understand intelligence. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Denis Cousineau et al. Confidence intervals in within-subject designs: A simpler solution to loftus and masson's method. *Tutorials in quantitative methods for psychology*, 1(1):42–45, 2005.

Logan Cross, Jeff Cockburn, Yisong Yue, and John P O'Doherty. Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments. *Neuron*, 109(4):724–738, 2021.

DeepSeek-AI. Deepseek-v3.2: Pushing the frontier of open large language models, 2025.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Mariya Hendriksen, Tabish Rashid, David Bignell, Raluca Georgescu, Abdelhak Lemkhenter, Katja Hofmann, Sam Devlin, and Sarah Parisot. Adapting vision-language models for evaluating world models. *arXiv preprint arXiv:2506.17967*, 2025.

Mark K Ho, David Abel, Thomas L Griffiths, and Michael L Littman. The value of abstraction. *Current opinion in behavioral sciences*, 29:111–116, 2019.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Patrick J Mineault, Thomas L Griffiths, and Sean Escola. Cognitive dark matter: Measuring what ai misses. *arXiv preprint arXiv:2603.03414*, 2026.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage*, 197:482–492, 2019.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Thomas Pouncy and Samuel J Gershman. Inductive biases in theory-based reinforcement learning. *Cognitive Psychology*, 138:101509, 2022.

Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114, 2018.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

Momchil S Tomov, Pedro A Tsividis, Thomas Pouncy, Joshua B Tenenbaum, and Samuel J Gershman. The neural architecture of theory-based reinforcement learning. *Neuron*, 111(8):1331–1344, 2023.

Pedro A Tsividis, Joao Loula, Jake Burga, Nathan Foss, Andres Campero, Thomas Pouncy, Samuel J Gershman, and Joshua B Tenenbaum. Human-level reinforcement learning through theory-based modeling, exploration, and planning. *arXiv preprint arXiv:2107.12544*, 2021.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the planning abilities of large language models-a critical investigation. *Advances in Neural Information Processing Systems*, 36:75993–76005, 2023.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Shengjie Wang, Shaohuai Liu, Weirui Ye, Jiacheng You, and Yang Gao. Efficientzero v2: Mastering discrete and continuous control with limited data. *arXiv preprint arXiv:2403.00564*, 2024.

Chang Yang, Xinrun Wang, Junzhe Jiang, Qinggang Zhang, and Xiao Huang. Evaluating world models with llm for decision making. *arXiv preprint arXiv:2411.08794*, 2024.

Lance Ying, Ryan Truong, Prafull Sharma, Kaiya Ivy Zhao, Nathan Cloos, Kelsey R Allen, Thomas L Griffiths, Katherine M Collins, José Hernández-Orallo, Phillip Isola, et al. Ai gamestore: Scalable, open-ended evaluation of machine general intelligence with human games. *arXiv preprint arXiv:2602.17594*, 2026.

A Use of LLMs in this Paper

Large Language Models were used as writing assistants for this paper and coding assistants for the experiments in this paper. In both cases, the final product were heavily edited by the authors.

B Encoding Performances for All Layers and Regions of Interest

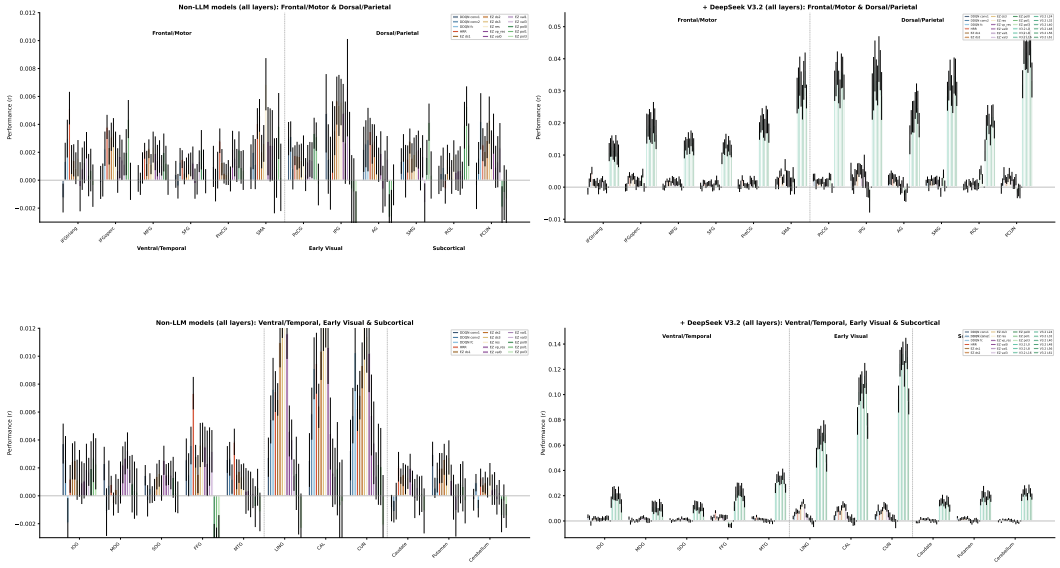

Figure 5: **Encoding performance across all model layers and ROIs.** Each bar represents a single model layer. Left panels: non-LLM models — DDQN (conv1, conv2, fc), HRR (EMPA theory embeddings), and EfficientZero (11 layers: 4 representation network, 3 value head, 3 policy head, 1 shared trunk). Right panels: same models with DeepSeek V3.2 (sampled every 8 layers; note y-axis rescaling). Top row: frontal/motor and dorsal/parietal ROIs. Bottom row: ventral/temporal, early visual, and subcortical ROIs. Error bars indicate ±1 CM-corrected SEM across subjects (N = 21). Dashed lines separate ROI groups.