COGNITIVE SCIENCE AND AI

Centaur: A Foundation model of Human Cognition

PROBLEM STATEMENT

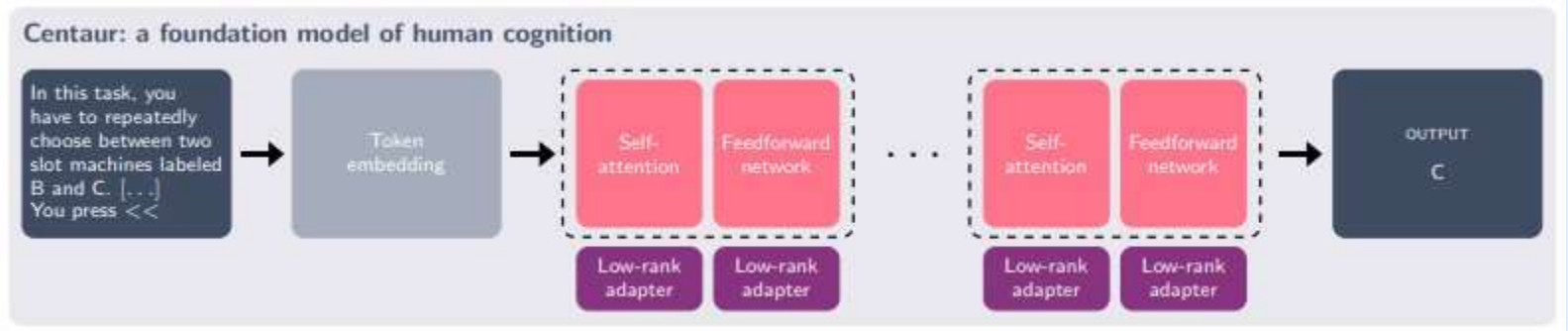
The Challenge of Modeling Human Cognition

- The human mind is **highly general**, capable of solving both **simple and complex problems**.
- **Current computational models** are **domain-specific**, excelling in narrow tasks (e.g., **AlphaGo** in Go) but failing to generalize.
- Cognitive models like **Prospect Theory** explain **decision-making** but not **learning, exploration, or reasoning**.
- Psychology aims to develop a **unified theory of cognition**, yet no model comprehensively captures human intelligence across domains.
- **Key Question:** *Can we build a general model that accurately predicts and simulates human behavior in diverse tasks?*

INTRODUCTION

Understanding Human Cognition & The Need for a Unified Model

- Humans effortlessly learn, reason causally, and act with curiosity, yet AI models remain narrow and task-specific.
- Cognitive science seeks domain-general models to capture the full range of human intelligence.
- Existing models, like AlphaGo or Prospect Theory, excel in specific areas but fail to generalize.
- Inspired by pioneers like **Newell (1990)**, who emphasized unified theories of cognition, Centaur takes a **data-driven approach** to bridge this gap.
- A unified model like Centaur could revolutionize our understanding of **decision-making, learning, and behavior prediction**.



What is Centaur?

Centaur is a **computational model** designed to predict and simulate **human behavior** in any experiment expressible in **natural language**. It is the **first real candidate for a unified theory of cognition**, overcoming the limitations of traditional **domain-specific cognitive models**.

How Centaur Works

- Fine-tuned on large data from the state-of-the-art language model Llama 3.1 70B to predict and simulate human behavior.
- Uses **low-rank adaptation layers** to efficiently specialize in cognitive tasks while maintaining general language capabilities.
- Accurately predicts human behavior even in new tasks, unseen scenarios, and modified problem structures, outperforming traditional cognitive models.
- After fine-tuning, its internal representations align closely with human neural activity, despite not being explicitly trained for it.

PSYCH 101 DATASET

- The **Psych-101 dataset** is a novel, large-scale data set covering trial-by-trial data from 160 psychological experiments
- Made by transcribing each of these 160 experiments into natural language, which provides a common format for expressing vastly different experimental paradigms.
- Provides a large corpus of human behavior to fine-tune Centaur in a data-driven manner containing over 10,000,000 human choices and including many canonical studies from various domains.

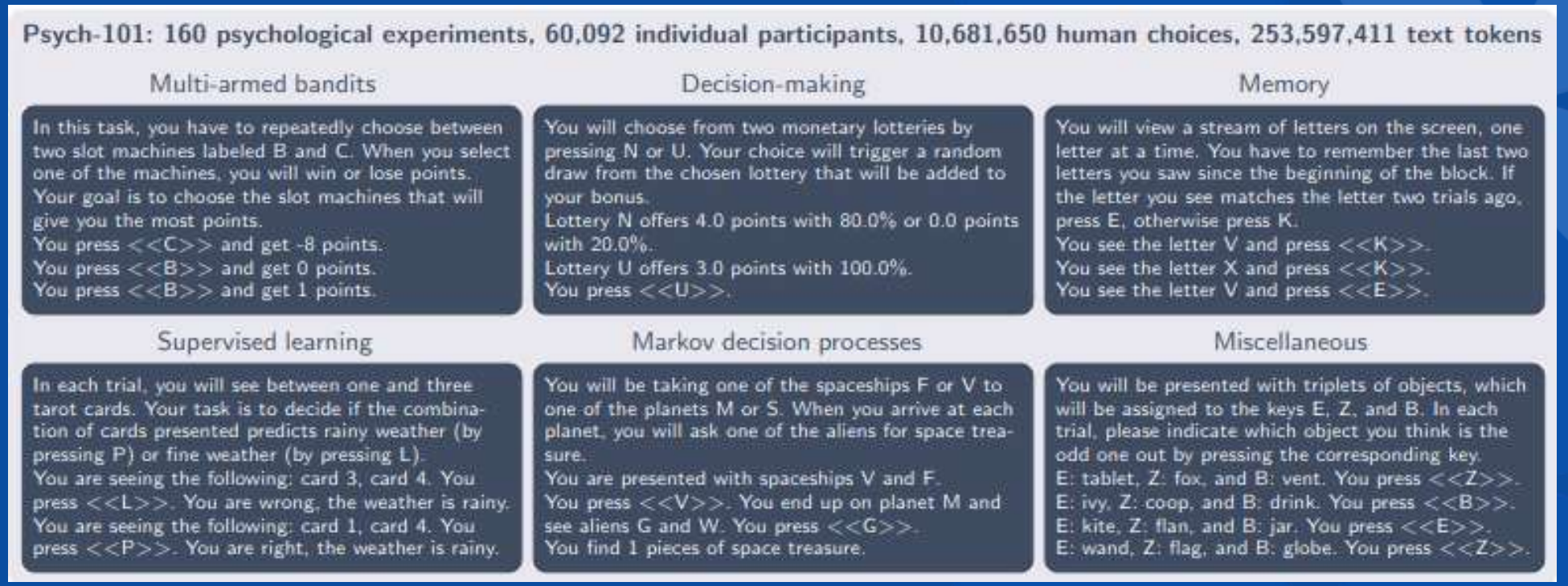
THE DOMAINS OF TASKS IN PSYCH-101

1. Multi-Armed Bandits
2. Decision-Making
3. Memory
4. Supervised Learning
5. Markov Decision Process
6. Miscellaneous

EXAMPLES OF THE TASKS

- Some examples of the tasks used to create dataset include:
  - Shephard Categorization
  - Drifting Four-Armed Bandit
  - Multiple Cue Judgement
  - Digit Span, etc.

Psych-101 Dataset



DOMAIN SPECIFIC COGNITIVE MODELS

- **14 cognitive and statistical models** were selected covering most Psych-101 experiments.
- Models were **trained on participant data** and evaluated using a **predictive pseudo-R^2 measure**.
- Out-of-distribution evaluations used the **most similar experiment** to fit parameters.
- Each model was implemented in **PyTorch** and optimized via **log-likelihood maximization**.

GENERALIZED CONTEXT MODEL

- Applied to **Shepard categorization, Medin categorization, and Weather Prediction Task**.
- Uses **similarity-based classification** to predict human responses.
- Core equation involves **exponential weighting** of past observations.

It uses the following log-likelihood:

$$p(c_t = i | x_t = \mathbf{x}_t) \propto \exp\left(\beta \sum_{k=1}^{t-1} \exp\left(-||\mathbf{x}_k - \mathbf{x}_t||_2\right) \cdot 1\left[y_k = i\right]\right)$$

where $\mathbf{x}_t$ are the features of the item observed at trial t and y_t is the corresponding class label. β is a free parameter of the model.

PROSPECT THEORY MODEL

- Applied to **CPC18, choices13k, and Decisions from Description**.
- Captures human **risk preferences and decision biases**.
- Uses **probability weighting functions** and a **nonlinear utility function**.

It uses the following log-likelihood:

$$p(c_t = i | p_i = \mathbf{p}_i, x_i = \mathbf{x}_i) \propto \exp\left(\exp(\beta)\left(\pi(\mathbf{p}_i)^\top u(\mathbf{x}_i)\right)\right)$$

$$\pi(\mathbf{p}_i) = \text{sigmoid}(a) + \text{sigmoid}(b) \quad \mathbf{p}_i$$

$$u(\mathbf{x}_i) = \begin{cases} \text{sigmoid}(c) \cdot \mathbf{x}_i^{\text{sigmoid}(d)} & \text{where } \mathbf{x}_i \geq 0 \\ -\text{sigmoid}(e) \quad (-\text{sigmoid}(f)\,\mathbf{x}_i)^{\text{sigmoid}(g)} & \text{where } \mathbf{x}_i < 0 \end{cases}$$

where $\mathbf{p}_i$ is the vector of probabilities and $\mathbf{x}_i$ is the vector of values for each possible outcome in option i. β, a, b, c, d, e, f, and g are free parameters of the model.

HYPERBOLIC DISCOUNTING MODEL

- Applied to **Intertemporal Choice** tasks.
- Models how humans **devalue future rewards** over time.
- Uses **nonlinear discounting functions** for reward valuation.

It uses the following log-likelihood:

$$p(c_t = i | x_i = x_i, \gamma_i = \gamma_i) \propto \exp\left(\beta\left(x_i \cdot \frac{1}{1 + (a \cdot \gamma_i)}\right)\right)$$

where x_i is the reward and γ_i is the delay of delivery for option i. β and a are free parameters of the model.

DUAL-SYSTEMS MODEL

- Applied to **Two-Step Task**.
- Balances **model-free (habit-based)** and **model-based (deliberative)** decision-making.
- Incorporates **Q-learning and reinforcement learning principles**.

It uses the following log-likelihood:

$$p(c_t = i|s_t = s) \propto \begin{cases} \exp\left(\beta\left(\text{sigmoid}\left(\tau\right) Q^{\text{MB}}_{s,i} + \left(1 - \text{sigmoid}\left(\tau\right)\right) Q^{\text{MF}}_{s,i}\right)\right) & \text{if } s = 0 \\ \exp\left(\beta Q^{\text{MF}}_{s,i}\right) & \text{if } s > 0 \end{cases}$$

where $Q^{\text{MB}}_{s,i}$ and $Q^{\text{MF}}_{s,i}$ are model-based and model-free value estimates that are computed as described in [21]. β and τ are free parameters of the model. We also included a

RESCORLA-WAGNER MODEL

- Applied to **Drifting Four-Armed Bandit, Horizon Task, Iowa Gambling, and more**.
- A **reinforcement learning** model that updates values based on **prediction errors**.

$$p(c_t = i) \propto \exp\left(aV_{i,t} + bS_{i,t} + cI_{i,t}\right)$$

$$V_{i,t} = \begin{cases} V_{i,t-1} + \text{sigmoid}\left(\alpha^+\right)\left(r_{t-1} - V_{i,t-1}\right) & \text{if } c_{t-1} = i \text{ and } r_{t-1} - V_{i,t-1} \geq 0 \\ V_{i,t-1} + \text{sigmoid}\left(\alpha^-\right)\left(r_{t-1} - V_{i,t-1}\right) & \text{if } c_{t-1} = i \text{ and } r_{t-1} - V_{i,t-1} < 0 \\ V_{i,t-1} & \text{otherwise} \end{cases}$$

$$S_{i,t} = 1\left[c_{t-1} = i\right]$$

$$I_{i,t} = \sum_{k=1}^{t-1} 1\left[c_k = i\right]$$

$$V_{i,1} = d$$

$$S_{i,1} = 0$$

$$I_{i,1} = 0$$

where r_t is the reward obtained in trial t. α^+, α^-, a, b, c, and d are free parameters of the model.

How much do I expect to gain?
How likely am I to repeat my last choice?
How often have I chosen this option?

RESCORLA-WAGNER MODEL WITH CONTEXT

- Applied to **Conditional Associative Learning**.
- Extends **Rescorla-Wagner Model** with **context-dependent learning**.

It uses the following log-likelihood:

$$p(c_t = i | s_t = s) \propto \exp\left(\beta V_{s,i,t}\right)$$

$$V_{s,i,t} = \begin{cases} V_{s,i,t-1} + \text{sigmoid}\left(\alpha\right)\left(r_{t-1} - V_{s,i,t-1}\right) & \text{if } c_{t-1} = i \text{ and } s_{t-1} = s \\ V_{s,i,t-1} & \text{otherwise} \end{cases}$$

$$V_{s,i,1} = d$$

where r_t is the reward obtained in trial t. α, β, and d are free parameters of the model.

LINEAR REGRESSION MODEL

- Applied to **Multiple-Cue Judgment and Gardening Task**.
- Predicts human choices using **weighted feature combinations**.

It uses the following log-likelihood for multiple-cue judgment:

$$p(c_t = i | x_t = \mathbf{x}_t) \propto \exp\left(\beta \left(\mathbf{w}_t^\top \mathbf{x}_t - i\right)^2 + \gamma\right)$$

It uses the following log-likelihood for the gardening task:

$$p(c_t = \text{accept} | x_t = \mathbf{x}_t) \propto \exp\left(\beta \mathbf{w}_t^\top \mathbf{x}_t\right)$$
$$p(c_t = \text{reject} | x_t = \mathbf{x}_t) \propto \exp(0)$$

and the following learning rule for both tasks:

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \alpha \left(r_{t-1} - \mathbf{w}_{t-1}^\top \mathbf{x}_{t-1}\right) \mathbf{x}_{t-1}$$
$$\mathbf{w}_1 = \mathbf{d}$$

where r_t is the reward obtained in trial t and $\mathbf{x}_t$ are the observed features. α, β, γ, and $\mathbf{d}$ are free parameters of the model.

WEIGHTED-ADDITIVE MODEL

- This model was used for the following experiments:
  - Multi-attribute decision-making
- It uses the following log-likelihood:
  - $p(c_t = i|x_i = \mathbf{x}_i) \propto \exp(\mathbf{w}^T\mathbf{x}_i)$
  - where $\mathbf{x}_i$ is the vector of features for option i and $\mathbf{w}$ are free parameters of the model.

DECISION-UPDATED REFERENCE POINT MODEL

- This model was used for the following experiments:
  - Columbia card task

It uses the following log-likelihood:

$$p(c_t = \text{sample} | x_{\text{win}}, x_{\text{loss}}, p_{\text{win}}, p_{\text{loss}}) \propto \exp\left(h \cdot \left(x_{\text{win}} \cdot p_{\text{win}} + x_{\text{loss}} \cdot p_{\text{loss}}\right) + i\right)$$

$$p(c_t = \text{stop} | x_{\text{win}}, x_{\text{loss}}, p_{\text{win}}, p_{\text{loss}}) \propto \exp\left(j\right)$$

$$\pi\left(p\right) = \text{sigmoid}\left(a\right) + \text{sigmoid}\left(b\right) \cdot p$$

$$u\left(v\right) = \begin{cases} \text{sigmoid}\left(c\right) \cdot v^{\text{sigmoid}(d)} & \text{where } v \geq 0 \\ -\text{sigmoid}\left(e\right) \quad \left(-\text{sigmoid}\left(f\right) v\right)^{\text{sigmoid}(g)} & \text{where } v < 0 \end{cases}$$

where x_{win} and x_{loss} are the values that can be won or lost respectively, and p_{win} and p_{loss} are the corresponding probabilities. a, b, c, d, e, f, g, h, i, and j are free parameters of the model.

ODD-ONE-OUT MODEL

- This model was used for the following experiments:
  - THINGS odd-one-out
- It uses the following log-likelihood:
  - $p(c_t = i|x_i, x_j , x_k) \propto \exp(\mathbf{x}_j^T \mathbf{x}_k)$
- where x_i, x_j , and x_k are the observed objects with their corresponding embeddings
- $\mathbf{x}_i$, $\mathbf{x}_j$, and $\mathbf{x}_k \in R^{16}$ that are free parameters of the model.

MULTI-TASK REINFORCEMENT LEARNING MODEL

- This model was used for the following experiments:
  - Multi-task reinforcement learning
  - Zoopermarket

GP-UCB MODEL

- This model was used for the following experiments:
  - Spatially correlated multi-armed bandit
  - Structured bandit
- It uses the following log-likelihood:
  - $p(c_t = i) \propto \exp(\beta(\mathbf{m}_{i,t} + \exp(\gamma)\,\mathbf{s}_{i,t}))$
- where $\mathbf{m}_{i,t}$ and $\mathbf{s}_{i,t}$ are obtained via Gaussian Process regression with a radial basis function kernel. β and γ are free parameters of the model.

RATIONAL MODEL

- This model was used for the following experiments:
  - Balloon analog risk task
  - N-back
  - Digit span
  - Go/no-go
  - Recent probes
  - Serial reaction time task
- It uses the following log-likelihood:
  - $p(c_t = i|o_t = j) \propto \exp(\Theta_{j,i})$
- where j is the optimal choice at trial t. $\Theta \in R^{Nc \times Nc}$ are free parameters of the model.

LOOKUP TABLE MODEL

- This model was used for the following experiments:
  - Grammar judgement
- It uses the following log-likelihood:
  - $p(c_t = i) \propto \exp(\boldsymbol{\Theta}_{t,i})$
- where $\boldsymbol{\Theta} \in R^{T \times Nc}$ are free parameters of the model.

INTRODUCTION TO LLAMA 3.1 70B

- **What is LLaMA?**
  - LLaMA (Large Language Model Meta AI) is an advanced open-source language model developed by Meta AI.
  - LLaMA 3.1 70B refers to a version with **70 billion parameters**, enabling complex language understanding and generation.
- **Key Features:**
  - State-of-the-art performance in language modeling tasks.
  - Pre-trained on vast datasets to capture diverse knowledge.
  - Open-source availability for research and development.
- **Why Use LLaMA?**
  - Robust performance for complex NLP tasks.
  - Flexibility for further customization through fine-tuning.
  - Efficient resource utilization despite large-scale parameters.

FINE-TUNING PROCEDURE (QLORA METHOD)

- **Base Model:**
  - LLaMA 3.1 70B model used as the foundation.
- **QLoRA Technique:**
  - **Quantized Low-Rank Adaptation**: Adds low-rank adapters to a 4-bit quantized base model.
  - Keeps the base model **frozen** while training only the new adapter parameters.
- **Fine-Tuning Process:**
  - **One Epoch**: Trained on the entire dataset to prevent overfitting.
  - **Loss Calculation**: Backpropagation applied **only to human responses**, masking other tokens.
- **Training Parameters:**
  - Effective batch size: 32
  - Learning rate: 0.00005
  - Weight decay: 0.01
  - Optimizer: 8-bit AdamW with 100-step linear warm-up

MODEL OVERVIEW (CENTAUR SYSTEM)

- **Built on LLaMA 3.1 70B:**
  - Utilizes LLaMA's extensive knowledge as a foundational backbone.
- **Customization via QLoRA:**
  - Low-rank adapters added to all non-embedding layers (Rank=8, Scale=8).
  - Additional parameters are only **0.15%** of the base model.
- **Training Details:**
  - Fine-tuned on **Psych-101 dataset**.
  - Masked loss for non-human responses to capture human behavior.
  - Took **~5 days** on an **A100 80GB GPU**.
- **Outcome:**
  - Efficient model adaptation without altering the core architecture.
  - Focused on human response modeling for improved behavioral understanding.

EVALUATION METRICS

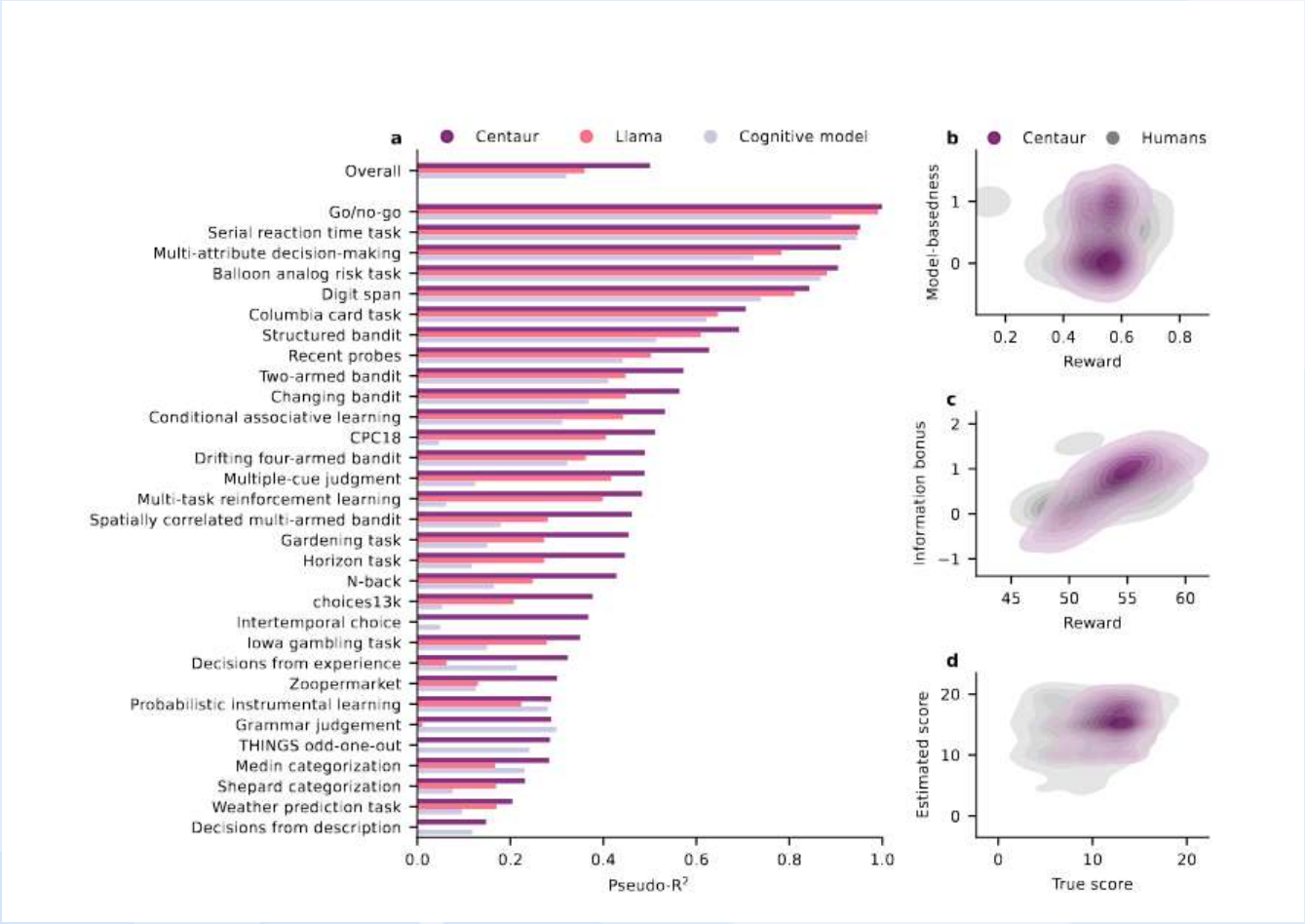
pseudo-R^2 measure is used to evaluate all the models

WHAT IS PREUDO-R^2?

- R^2 measures how well the model explains the variance of the target variable or how well the **predictions** align with the **ground truth**, **aka it indicates the quality of fit of the model's predictions to the observed data**; a measure of prediction accuracy.
- Pseudo-R^2 approximates this concept for models with discrete target variables, providing an analogous measure of model fit that takes the average log-likelihood of human responses for a given model and normalizes it using the average log-likelihood of a model that guesses responses uniformly.

$$\mathrm{R}^2 = 1 - \frac{\log p_{\text{model}}(\mathcal{D})}{\log p_{\text{guess}}(\mathcal{D})}$$

PERFORMANCE ON PSYCH-101.



a, Pseudo-R2 values for different models across experiments
b, Model simulations on the two-step task
c, Model simulations on the horizon task.
d, Model simulations on a grammar judgement
Task.

- The average improvement across experiments after finetuning was 0.14 (Centaur pseudo-R2 = 0.50; Llama pseudo-R2 = 0.36).
- The average improvement in predicting human behavior over the domain-specific cognitive models was 0.18(Centaur pseudo-R2 = 0.50; cognitive models pseudo-R2 = 0.32).

PROBING INCREASINGLY COMPLEX GENERALIZATION ABILITIES

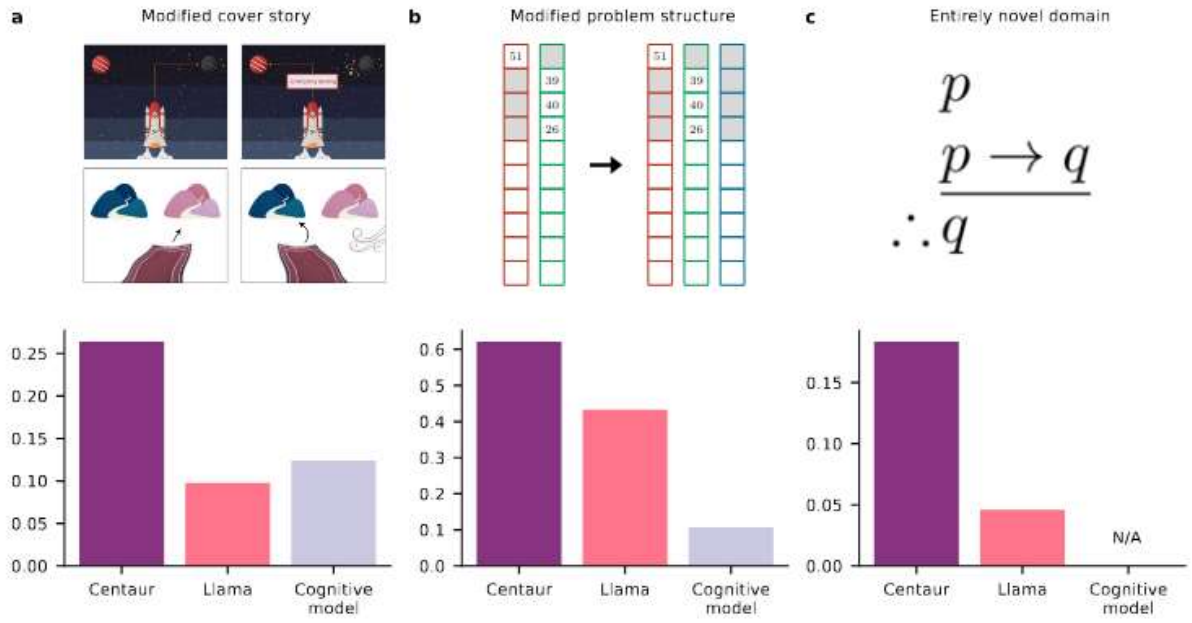


Fig. 3 Evaluation in different held-out settings. **a**, Pseudo-R^2 values for the two-step task with a modified cover story [24]. **b**, Pseudo-R^2 values for a three-armed bandit experiment [25]. **c**, Pseudo-R^2 values for an experiment probing logical reasoning [26]. Centaur outperforms both Llama and domain-specific cognitive models when faced with modified cover stories, problem structures, and entirely novel domains.

INTERNAL REPRESENTATIONS BECOME MORE ALIGNED TO HUMAN NEURAL ACTIVITY

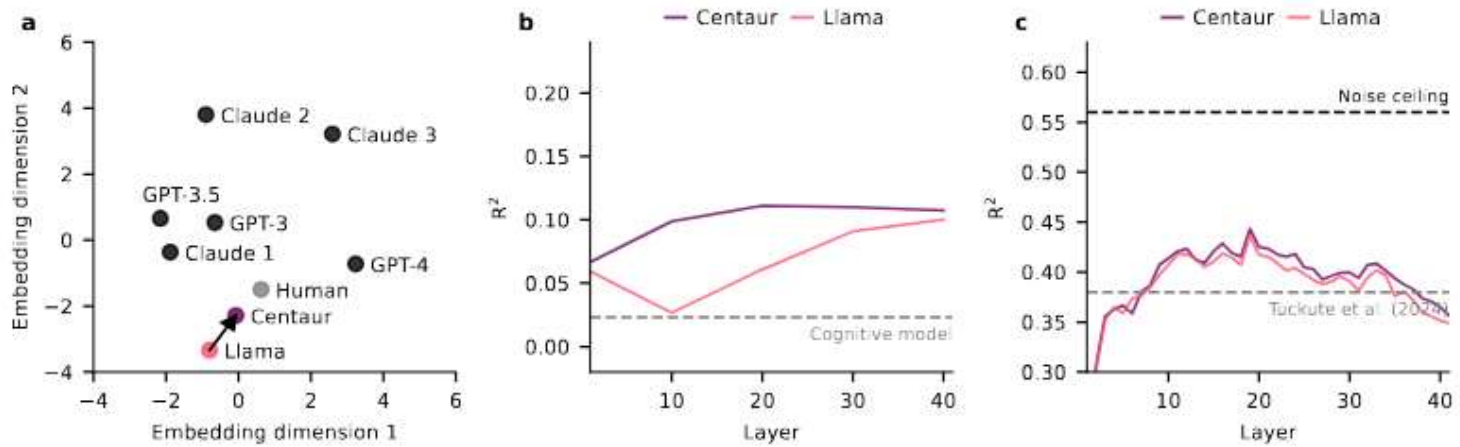
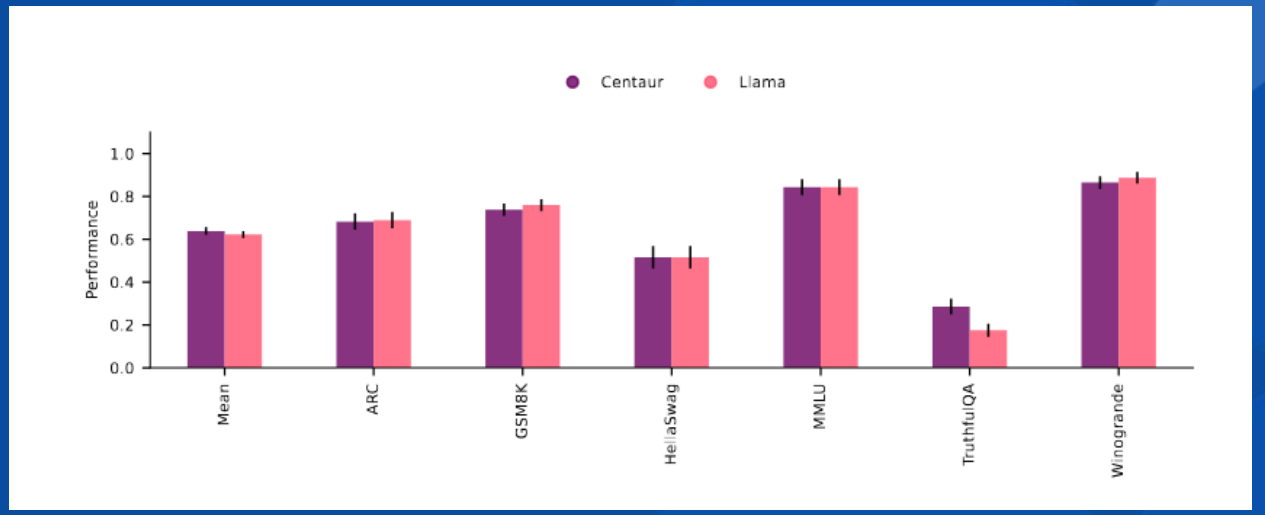


Fig. 4 Alignment between humans and Centaur. **a**, Multidimensional scaling embedding of the ten behavioral metrics in CogBench [29] for different models. **b**, R^2 values indicating how well human neural activity in the two-step task [33] can be decoded using Centaur's internal representations extracted from different layers. **c**, R^2 values indicating how well human neural activity in a sentence-reading task [34] can be decoded using Centaur's internal representations extracted from different layers.

BENCHMARKS

metabench

- a sparse benchmark containing several canonical benchmarks from the machine learning literature.
- Centaur maintains the level of performance of Llama, indicating that finetuning on human behavior did not lead to deterioration in other tasks.
- Performance on TruthfulQA– which measures how models mimic human falsehoods – even improved significantly with finetuning.



BENCHMARKS

cogbench

- includes ten behavioral metrics derived from seven cognitive psychology experiments.
- relative to Llama, Centaur's performance improves in all experiments.
- Centaur becomes more similar to human subjects in all ten behavioral metrics

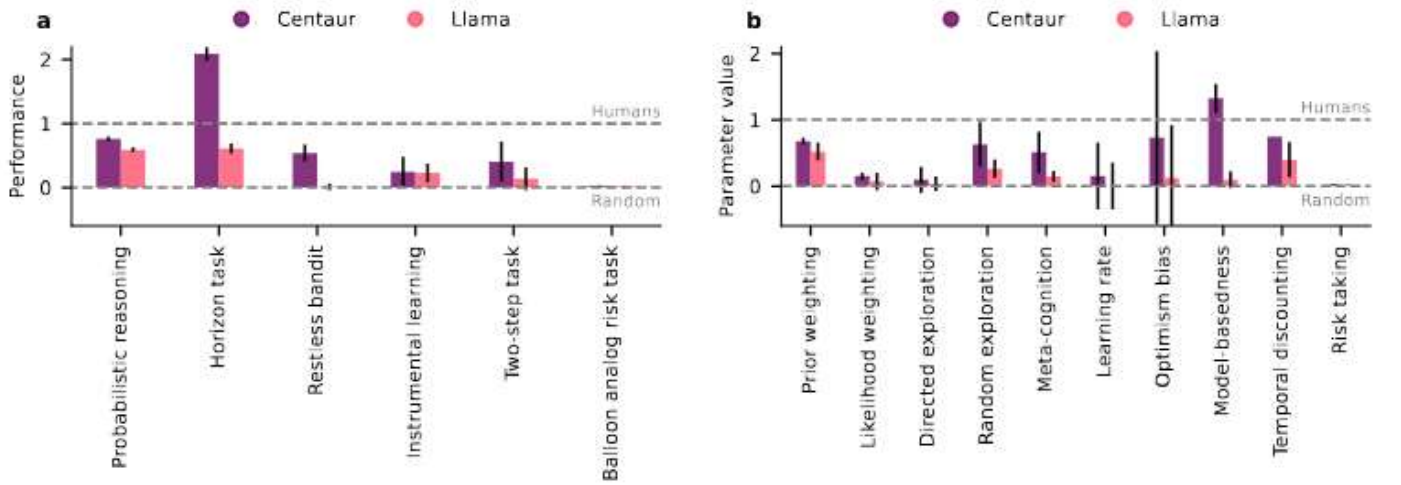


Fig. 7 CogBench [29] results. **a**, Performance-based metrics. **b**, Behavioral metrics. All metrics are human-normalized: a value of zero corresponds to a random agent, while a value of one corresponds to the average human subject.

NEURAL ALIGNMENT ANALYSIS (TWO-STEP TASK)

Model Used:

- Regularized **linear regression** to predict **fMRI** data from **Centaur** and **LLaMA** internal representations.
- Separate models for each **participant** and **region**.

Procedure:

- **Train-Test Split:** Fitted on **two scanning blocks**, evaluated on the **third**.
- **Cross-Validation:** Regularization strength chosen via **nested cross-validation**.

Region of Interest (ROI) Handling:

- Used **Harvard-Oxford atlas** to split beta maps into **cortical** and **subcortical** ROIs.
- Averaged betas within each ROI to reduce voxel complexity.

Metric:

- Reported **R^2** values are averaged across **all ROIs**.

DATA PROCESSING & MODEL ALIGNMENT

Internal Representation Extraction:

- Extracted from **models' residual stream**.
- Applied **PCA** to retain components explaining **95% variance**.

fMRI Preprocessing (fMRIPrep 24.0):

- Scans aligned to **MNI152NLin2009cAsym** atlas.
- Default **fMRIPrep** settings used.

General Linear Models (GLM) Setup:

- Separate **GLMs** for each subtrial (e.g., second step, feedback).
- Included noise regressors: 6 **motion parameters** and **framewise displacement**.

NEWELL TEST

A set of criteria proposed by cognitive scientist Allen Newell to evaluate the comprehensiveness and robustness of theories in cognitive science.

Centaur has demonstrated compliance with several of Newell's criteria.

NEWELL TESTS PASSED

- **Behave as an (almost) arbitrary function of the environment**
  - The most important criterion according to Newell. Centaur fulfills it more than any previous model. Yet, scope still limited to experiments that can be expressed in natural language. Will be an important avenue for future research to transfer ability to real-world applications.
- **Operate in real time**
  - Centaur can simulate human behavior in (almost) real-time. For example, running open-loop simulation of a typical two-step task experiment takes around 30 minutes (takes around 20 minutes for average human participant). Can be further optimized.

- **Exhibit rational, that is, effective adaptive behavior**
  - Bayesian inference is the gold standard for rational and adaptive behavior. Previous work has shown that systems that engage in in-context learning (which Centaur uses) implement Bayesian inference implicitly.

- **Use vast amounts of knowledge about the environment**
  - Large language models are the biggest knowledge bases we have to date. As Centaur is built on top of a state-of-the-art language model, it fulfills this criterion by design.

- **Behave robustly in the face of error, the unexpected, and the unknown**
  - Extensive out-of-distribution evaluations clearly demonstrate that Centaur has this ability.

- **Integrate diverse knowledge**
  - This was originally a criterion on symbols and abstractions. At the basic level, Centaur is a system that processes language. Language is a symbolic system, meaning that Centaur fulfills this criterion

- **Use Natural Language:**
  - Interpret and generate human language fluently.
  - Built on a language model, enabling seamless processing and response in natural language.
- **Exhibit Self-Awareness and a Sense of Self:**
  - Maintain internal representations of itself or the entity producing behavior.
  - Infers user identity from behavioral patterns despite being trained on population-level data.
- **Learn from the Environment:**
  - Update knowledge and behavior based on new environmental input.
  - Models human learning processes across experiments requiring environmental adaptation.

LIMITATIONS OF CENTAUR IN THE NEWELL TEST

Acquire Capabilities Through Development:

- Centaur does not model how cognitive abilities emerge over time.

Arise Through Evolution:

- No claims about the evolutionary processes underlying human cognition.

Be Realizable Within the Brain:

- Centaur's internal representations align with human neural activity, but differences remain between transformer models and brain architecture.

Thank You

PRESENTED BY:

- BHUMIKA JOSHI, 2022121006
- VISHNA PANYALA, 2021101044