Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Taming the Centaur(s) with LAPITHS: a framework for a theoretically grounded interpretation of AI performances

Matteo Da Pelo
University of Cagliari https://orcid.org/0009-0006-5875-726X
Alessio Donvito
University of Bari https://orcid.org/0009-0007-7546-0342
Claudio Frongia
University of Cagliari https://orcid.org/0009-0005-4305-2278
Pietro Salis
University of Cagliari https://orcid.org/0000-0001-6973-1440
Antonio Lieto
alieto@unisa.it
University of Salerno https://orcid.org/0000-0002-8323-8764

Research Article

Keywords: Computational Cognitive Science, Cognitive Modeling, Artificial Intelligence, Cognitive Plausibility, Unified Theory of Cognition

Posted Date: May 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9575224/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Taming the Centaur(s) with LAPITHS: a framework for a theoretically grounded interpretation of AI performances

Matteo Da Pelo[1], Alessio Donvito[2,4], Claudio Frongia[3], Pietro Salis[1], Antonio Lieto[4,5]

[1]Department of Education, Psychology, Philosophy, University of Cagliari, Cagliari, Italy.
[2]DIRIUM, University of Bari, Bari, Italy.
[3]Department of Electrical and Electronics Engineering, University of Cagliari, Cagliari, Italy.
[4]Cognition Interaction and Intelligent Technologies Laboratory, University of Salerno, Salerno, Italy.
[5]ICAR-CNR, Palermo, Italy.

Abstract

We introduce a framework called LAPITHS (Language-model Analysis through Paradigm-grounded Interpretations of Theses about Human-likenesS) and use it to show that several major claims advanced by models such as CENTAUR—proposed as an artificial Unified Model of Cognition—are not theoretically or empirically justified. LAPITHS provides a principled reference point for counteracting the current behaviouristic tendency in AI research to interpret the human-level performances of transformer-based language models as evidence of human-like underlying computation and, by extension, as signs of cognitive abilities. The novelty of LAPITHS lies in making explicit the arguments grounded in two quantitative assessments: (i) the Minimal Cognitive Grid, a theoretically motivated method for estimating the cognitive plausibility of artificial systems, and (ii) a behavioural comparison showing that results similar to those reported for CENTAUR-like models can be reproduced by other systems that do not satisfy the structural constraints typically associated with cognitive plausibility, and whose outputs do not provide independent explanatory insight into human cognition.

Keywords: Computational Cognitive Science, Cognitive Modeling, Artificial Intelligence, Cognitive Plausibility, Unified Theory of Cognition

1 Introduction

In recent years, the rapid development of large-scale transformer-based systems has significantly reshaped the landscape of both artificial intelligence and cognitive science (Sartori and Orrù 2023; Qu and Wang 2024; Tuckute et al. 2024; Nie et al. 2025; Wulff and Mata 2026; Radford et al. 2019; Touvron et al. 2023). Within this context, a growing line of research has advanced the idea that sufficiently large and properly trained language models may serve not merely as tools for prediction, but as candidates for modelling human cognition itself (Binz and Schulz 2023; Chen et al. 2026). A paradigmatic example of this trend is Centaur, a system built through the fine-tuning of a large language model (Llama 3.1 70B) on an extensive corpus of human behavioural data, namely Psych-101, comprising over ten million decisions collected across a wide range of psychological experiments. The model is presented as a general-purpose predictor of human behaviour across multiple domains, and has been associated with the broader ambition of contributing to a unified, data-driven theory of cognition (Binz et al. 2025). Despite these promising developments, a number of conceptual and methodological issues arise when one examines more closely the nature of the claims involved. In particular, the success of Centaur in predicting human responses across heterogeneous tasks raises a central problem: whether human-level behavioural predictive performance can legitimately be taken as evidence that the model implements cognitively plausible mechanisms. The model's design, grounded in large-scale fine-tuning on curated behavioural data, encourages the reproduction of human-level outputs. As such, its success may be understood as a consequence of optimisation toward behavioural agreement, rather than as evidence of shared underlying mechanisms. The roots of this problem can be traced to a broader tendency within contemporary AI research, namely the implicit adoption of a functionalist perspective in which matching input–output behaviour is treated as sufficient for establishing cognitive equivalence. As discussed in the literature, and explicitly addressed in this work, such a stance risks conflating behavioural adequacy with cognitive plausibility (Quattrociocchi et al. 2026; Palminteri and Wu 2026; Lerchner 2026). In order to address this problem, the present work introduces a theoretically grounded framework—LAPITHS (Language models Analysis through Paradigm-grounded Interpretations of Theses about Human-likeness)—designed to disentangle performance from cognitive interpretation.[1] The framework combines a conceptual analysis of the conditions under which cognitive ascriptions are warranted with an operational evaluation grounded in the Minimal Cognitive Grid (MCG) (Lieto 2021). The MCG provides a structured methodology for assessing artificial systems along three key dimensions: functional/structural ratio, generality, and performance match. Crucially, it treats behavioural success as just one component of a broader evidential framework, thereby allowing

[1] Lapiths was also the name of an ancient population fighting the Centaurs in the epic narration of Centauromachy, sculpted in the Parthenon.

for a more detailed assessment of cognitive plausibility. The empirical component of this work applies the LAPITHS framework to a series of behavioural and representational analyses centred on the two-step task, a canonical paradigm in reinforcement learning research (Daw et al. 2011) and a crucial task for assessing claims made about the model (since its performances have led to conclusions about both neural and behavioural alignment by the authors, see Section 5.3 for details). The results indicate that other state-of-the-art language models, not specifically trained on the task to handle but only informed via RAG about the structure of the task, are capable of achieving levels of behavioural fit comparable to those reported for Centaur, as measured by negative log-likelihood, with differences that are in some cases statistically not significant. Moreover, representational analyses based on fMRI data suggest that high levels of alignment with human neural activity can be reproduced by models that have not undergone task-specific fine-tuning. Taken together, these results support a more cautious interpretation of recent developments. While models such as Centaur demonstrate some strong predictive capabilities, their performance appears to reflect functional adequacy at the level of input–output behaviour rather than the implementation of cognitively plausible mechanisms (Quattrociocchi et al. 2025). The LAPITHS framework, and in particular the MCG analysis, makes explicit the extent to which such systems diverge from the structural constraints identified in cognitive theory, thereby providing a principled basis for evaluating their explanatory status. The structure of the paper is organised as follows. Section 2 introduces Centaur in detail, outlining its architecture, training methodology, and the composition of the Psych-101 dataset. Section 3 develops the theoretical foundations of the critique, focusing on the distinction between functional and structural models, the role of functionalism in AI, and the formulation of the ascription fallacy. Section 4 presents the LAPITHS framework, including the formalisation of the MCG and its application to the evaluation of artificial systems. Section 5 reports the empirical analyses, including behavioural replication experiments on the two-step task, statistical comparisons with baseline models, and an fMRI-based representational study. Finally, Section 6 discusses the broader implications of the findings for the evaluation of AI systems and for the project of cognitive modelling, highlighting directions for future research.

2 What Is Centaur

Centaur is a foundation Artificial Intelligence model based on Llama 3.1 70B (Bommasani et al. 2021; Touvron et al. 2023; Grattafiori et al. 2024). In the author's view, a crucial step toward a unified theory of cognition is the development of a computational model capable of predicting and simulating human behaviour across all domains (Wu et al. 2018). Based on the reported data, the authors considered Centaur as a candidate for a unified theory of cognition (Newell 1973, 1994; Anderson and Lebiere 2003; Laird et al. 2017) proposing the idea that such behavioural replication is an indicator of some underlying cognitive alignment between the model and the mechanisms of human cognition. Centaur, in fact, appears to effectively predict the decisions made by some human participants in the Psych 101 test dataset. The dataset was constructed by transcribing data from 160 psychological experiments into natural language. Each

prompt was designed to include the entire sequence of a complete session for a single participant in each trial. The transcription of each experiment was manually performed by the authors. The maximum length of each prompt was set at approximately 32,768 tokens. The dataset includes six categories of experiments, consisting of experiments on: (i) decision-making, (ii) memory, (iii) supervised learning, (iv) Markov decision processes, (v) multi-armed bandits, (vi) miscellaneous. After having built such dataset, Centaur was built in a data-driven manner through the fine-tuning of Llama specifically on Psych 101. Fine-tuning is a technique for adapting a model's initial dataset to a specific knowledge domain. This process also involves backpropagation and weight optimization, but it operates on a significantly smaller dataset compared to pre-training. Fine-tuning enables the model to achieve higher performance on task-specific applications. The fine-tuning technique employed for Centaur is QLoRA, or Quantised Low-Rank Adaptation. QLoRA preserves the base model's parameters, thus trying to freeze as much as possible the representations obtained during the training phase, while introducing low-rank adapters containing a small number of additional trainable parameters. This approach enhances the well-known LoRA method, achieving comparable performance while significantly reducing energy and memory consumption. This efficiency improvement appears to be QLoRA's primary contribution to the state of the art (Hu et al. 2021; Dettmers et al. 2023). As a fine-tuned system, Centaur performs remarkably well on the "cognitive decathlon", a benchmark introduced by the authors to evaluate the capacity of computational models to predict human behaviour. One of the most striking aspects of the model is that its internal representations appear to become increasingly aligned with human behavioural patterns during training. In other words, the model's internal representations seem to synchronize with patterns extracted from the behavioural data.

Despite these noteworthy achievements, several theoretical concerns arise regarding Centaur. In particular, some aspects of the model's design raise questions about its suitability as a robust cognitive predictor. Moreover, certain experimental findings suggest that the reported results may be replicable with other approaches without any explicit commitment to explanatory adequacy. Our analysis suggests that evaluating the performance of a fine-tuned large language model in experimental contexts requires careful examination of the training dataset itself. The dataset may contain implicit properties or latent regularities that contribute substantially to the model's apparent success. Consequently, understanding how the training data are constructed and organized becomes crucial for explaining how the model achieves its performance and for assessing the extent to which its results genuinely reflect cognitive modeling rather than dataset-specific learning effects (Gluck and Laird 2018). In the following paragraph, a fundamental methodological distinction in AI design is introduced, which underpins one of our central objections to the results and overall design of Centaur.

3 Functional vs. Structural Artificial Models

An important methodological distinction in AI and computational cognitive modeling is that between functional and structural models (or, more precisely, between functionally and structurally designed models). This distinction is central to the debate

in the explanatory role played by artificial models with respect to the natural cognitive systems that they take as sources of inspiration. Functionalism was introduced into the philosophy of mind by Hilary Putnam, beginning with his influential paper "Minds and Machines" (Putnam 1960), as an alternative to type-identity theory based on the idea that mental states are individuated by their functional roles rather than by their physical realisations. In particular, two tokens are assumed to belong to the same mental state if they stand in the same functional relations to other mental states and to the system's inputs and outputs. Functionalism led, in AI, to the definition of a design approach based on the notion of "functional equivalence" between some cognitive faculties (to be modelled) and the corresponding mechanisms implemented in AI programs (Lieto 2021). Indeed, its more radical formulation postulated the sufficiency, from an epistemological perspective, of a weak equivalence between cognitive processes and AI procedures, positing that, from an explanatory point of view, the relation between "natural mind" and "artificial software" could have been based purely on a macroscopic input-output equivalence. In the last decades, this position has been widely criticised. In particular, models and systems designed according to the "functionalist" perspective have been deemed inadequate for advancing the science of cognitive AI. As in the case of an airplane—which flies through mechanisms entirely different from those of a bird (Russell and Norvig 2012)—the mechanisms and design choices underlying such artefacts prevent them from playing a genuine explanatory role with respect to their analogous natural systems. This applies, for example, to all the main contemporary AI systems, including so-called "cognitive computing" technologies such as IBM Watson (Ferrucci et al. 2010), AlphaGo (Silver et al. 2016), and large language models like GPT-5 or DeepSeek R1 (Guo et al. 2025a,?). Despite the propaganda, such systems cannot be qualified as "cognitive" under a structurally grounded notion of cognition, insofar as they lack explanatory relevance with respect to: (i) how humans organise, retrieve, and reason over stored information when answering questions, or (ii) how humans plan and make decisions, etc. In other words, most current AI systems are "functionalists": they function as if they were natural cognitive systems in terms of output, but the internal mechanisms generating such outputs differ from those employed by humans. Therefore, artificial imitation of cognitive capacities does not imply implementation of the same underlying principles. As mentioned, this is an important aspect to point out, as there is a persistent tendency to attribute cognitive explanations to purely functional systems, a tendency reinforced by the loose and often misleading use of terms such as "cognitive computing", which is frequently employed as a broad label for systems capable of interacting with humans. In contrast to functionalism, a structural approach to AI design calls for a more strongly constrained correspondence between artificial systems (their internal architectures and processes) and their natural counterparts. According to this view, structurally constrained artificial models and systems can be useful both to advance the science of AI in terms of technological achievements and to play the role of "computational experiments", able to provide insights and results useful in refining or rethinking theoretical aspects concerning the target natural system used as a source of inspiration (Milkowski 2013; Cordeschi 2002). A problem arising in this

view is that it is not possible to build a completely structural and constrained artificial model, since an exact replica of a natural system is unattainable. The pursuit of increasingly structural models leads to an asymptotic regress to the microscopic physical world, culminating in the well-known Wiener paradox: "The best material model of a cat is another, or preferably the same, cat" (Rosenblueth and Wiener 1945). This paradox highlights the necessity of constructing proxy models rather than replicas. Similarly, Zenon Pylyshyn (Pylyshyn 1979) noted that: "if we do not formulate any restriction about the model, we obtain the functionalism of the Turing machine, and if we apply all the possible restrictions, we reproduce a whole human being". The key issue, therefore, is identifying the appropriate level of description and enforcing constraints at that level, in order to achieve human-like computation (Marr 1982). In this view, the functionalism/structuralism dichotomy is better understood as a continuum. Between the explanatory inadequacy of pure functional artificial models (which may nonetheless achieve impressive performance in specific tasks) and the impracticality of pure structural models, there exists a spectrum of plausible artificial proxy models with varying degrees of explanatory power. Having clarified these preliminary points, we may now turn to the central focus of the article: the problems that emerge from a close reading of the scientific presentation of Centaur. A critical examination reveals a number of issues, some of which are not unique to this work but instead reflect broader trends in the literature, extending beyond the specific contribution of Binz and colleagues.

3.1 Ascription Fallacy

A central problem is the so-called "ascription fallacy": i.e., conflating human-level behavioural performance with human-like computational mechanisms, and then explaining AI outputs using the same biological or cognitive theories we use to explain analogous human outputs (Quattrociocchi et al. 2025). Centaur is a salient case where this conflation can occur, for three connected reasons. First, Centaur's training objective structurally encourages behavioural isomorphism, not mechanistic homology (Orr et al. 2025). Centaur is trained to reproduce or predict human trial-by-trial responses in psychological tasks, via standard language-model fine-tuning (parameter-efficient adaptation on a pre-trained LLM). As a result, strong behavioural fit is unsurprising, since the optimisation target is explicitly defined in terms of agreement with observed human responses. However, fit-to-behaviour alone is not evidence that the system implements the same competence in a biological or cognitive sense. It is evidence of functional adequacy at the interface, namely that a mapping from task description and history to response can be learned. In the "ascription fallacy" framing, the mistake occurs when this kind of functional match is treated as if it licensed a structural (mechanistic or biological) attribution.

In particular, this happens when the competence of a given cognitive system for a particular task is assigned also to an artificial system for the mere equivalence of the obtained results, despite the structural differences of the mechanisms used by the natural and artificial systems to perform that task. Such ascription leads to the interpretation of the results of the artificial system in terms of the competence explaining the behaviour in the natural system (see Figure 1).

Second, the inference from accurate behavioural prediction to the possession of cognition is underdetermined. Centaur's headline contribution is broad behavioural prediction across many experiments and generalisation to certain distribution shifts. However, behavioural equivalence classes are large: many distinct internal organisations can implement the same input–output mapping, or match the same benchmark statistics, without sharing the causal or mechanistic properties that make a cognitive capacity in humans. This is exactly the "behaviourist trap" that the ascription fallacy critique targets (Li et al. 2026). Once evaluation privileges outputs alone, systems that exploit statistical regularities, retrieval, or shallow heuristics can be misread as instantiating the same competence humans deploy (Quattrociocchi et al. 2025). Centaur's success, therefore, establishes, at most, that a language-model-based architecture can serve as a high-coverage behavioural emulator for tasks encoded in natural language, not that it realises the same competence type as biological cognition. Third, "neural alignment" results are vulnerable to a reverse-inference leap. The Centaur paper reports that internal representations become "more aligned with human neural activity" after fine-tuning (Binz et al. 2025, p. 5). This type of result is often rhetorically taken to support a mechanistic reading, suggesting that the model is becoming brain-like. Methodologically, however, representational similarity is correlational and level-ambiguous. It can arise because the model and humans are exposed to similar task structures, stimulus descriptions, and response demands, without implying shared computational primitives, learning dynamics, or causal decompositions into human cognitive mechanisms. The ascription fallacy is triggered if one moves from "some representational alignment under a given analysis" to "therefore, the system instantiates the same cognitive process", especially when this move is used to explain human cognition in terms of the model's learned internal features.

3.2 Centaur is functionalist by default

The proponents of Centaur, viewing it as a promising candidate for a foundational model of cognition, underscore the importance of its performance aligning with a broad spectrum of human results (i.e., Psych-101). This raises an important consideration. The comparison between Centaur's outcomes and human results in the dataset predominantly focuses on matching performance outputs. Consequently, a functionalist perspective on these models and results is implicitly assumed, emphasising behavioural outputs. However, the function performed by a system S is distinct from the ways in which that function is implemented. Two systems, S and S_1, may realise the same cognitive function F while relying on entirely different architectures and mechanisms. Consequently, a genuinely cognitive system and a non-cognitive one may be evaluated as equivalent if assessment is based solely on behavioural outputs, regardless of whether the process involves actual cognition. For a more robust evaluation, artificial systems should be assessed along two dimensions: (1) functional design, concerning the performance of S in a task T, and (2) structural design, addressing the architecture and mechanisms that realise that performance T (Lieto 2021, 2022). This alternative perspective has broader implications, such as the adoption of a minimal cognitive grid (MCG), comprising a basic set of functional and structural criteria to assess the cognitive profile of artificial systems (Webb 2001; Lieto 2021, 2022). By

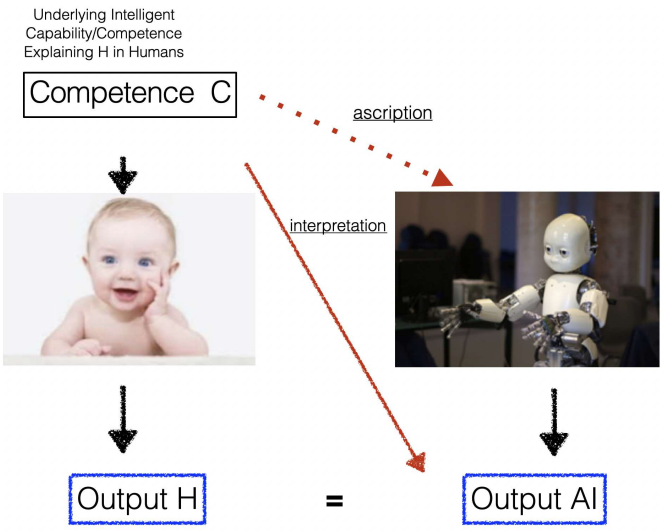


Fig. 1 A pictorial representation of the ascription fallacy problem, where the fact that the Humans and AI produce the same output, despite using structurally different mechanisms, leads to the unjustified ascription, to the AI system, of the theory of competence explaining the behaviour in the natural system.

contrast, Centaur's proponents appear to assume functionalism as the default stance without explicitly justifying it. Is this assumption warranted? Can we legitimately infer the human-likeness of systems such as Centaur on the basis of behavioural similarity alone? Perhaps this stance hinges on the shift from traditional symbolic artificial systems to neural networks. The argument suggests that this transition addresses the issue of structuralism at the outset, given the alleged resemblance between neural networks and the human brain. However, such resemblance is superficial. Contemporary neural networks differ in critical ways from human neural architectures. Contemporary neural networks differ in fundamental ways from biological neural systems. Mechanisms such as backpropagation, attention, and memory operate according to principles that diverge significantly from those observed in the brain (Crick 1989; Lillicrap et al. 2020). Thus, merely assuming that structuralism has been addressed through the shift from symbolic AI to neural networks is insufficient (Orr et al. 2025; Quattrociocchi et al. 2025). We are once again faced with the necessity of critically examining the functionalist assumptions advocated by Centaur's proponents. If its empirical results are to support ambitious claims about cognition, those assumptions must be explicitly defended. This simplistic functionalist assumption closely resembles a radical version of functionalism akin to behaviourism, where only behavioural outputs matter. As we will show in the following sections, this assumption is problematic.

4 The Framework

This section specifies a two-part methodological framework designed to test—and, where warranted, reject—the inference from linguistic output quality to possession

of a cognitive competence in the human/biological sense. The motivating concern is underdetermination: many heterogeneous systems can generate the same (or indistinguishably similar) input–output behaviour yet differ radically in internal organisation, causal structure, and explanatory relevance to human cognition. The framework treats output-based success as just *one component* of an evidential package, and it makes explicit what kind of additional evidence is required for a justified competence attribution. The framework has two tightly coupled components: (1) a theoretical component: a principled argument that output evidence is not identifiable with cognitive competence, plus an operational analysis of what counts as stronger evidence using a formalisation of the Minimal Cognitive Grid (MCG); (2) an experimental component (behavioural equivalence demonstrations): empirical constructions showing that the same outputs (or the same benchmark-level performance) can be produced by non-cognitive systems, and that this fact alone does not warrant cognitive ascription to those systems, revealing a flaw in output-only inference.

4.1 Minimal Cognitive Grid

To address the long-standing challenge of providing computational explanations of cognition (Milkowski 2013), the Minimal Cognitive Grid (MCG) (Lieto 2021) was introduced as a framework for evaluating the epistemological and explanatory status of artificial systems. Its primary aim is to assess the extent to which such systems align with relevant cognitive theories of the modelled phenomenon. Building on Webb's seminal proposal for characterising bio-inspired models (Webb 2001), and in line with the principles of Psychometric Artificial Intelligence (Bringsjord and Schimanski 2003; Bringsjord 2011), Lieto developed a model-agnostic evaluation approach grounded in the following dimensions: (1) Functional/Structural Ratio: this dimension concerns the analysis of the balance between functional and structural design components in the model. In complex artificial systems, it is possible to represent certain components in a purely functional manner, especially when their internal mechanisms are not central to the system's intended explanatory or pragmatic goal. Conversely, other components may require a structurally faithful modelling approach; (2) Generality: it captures the extent to which a system is capable of addressing diverse cognitive domains. A system that covers a broader range of tasks is considered more cognitively plausible, as it better reflects the versatility typical of human cognition; (3) Performance Match: this dimension requires a direct comparison between artificial and biological/cognitive systems, focusing on their performance across specific tasks through the use of standardised and validated benchmarks to assess the degree of alignment between artificial behaviour and human cognition. Two additional criteria are proposed to refine the evaluation: (a) error pattern analysis: the system should exhibit error types and distributions comparable to those observed in human performance; (b) execution time: the time taken by the artificial system to complete tasks should approximate human timing under analogous conditions. Taken together, the three dimensions of the Minimal Cognitive Grid provide a non-subjective method for assessing the degree of cognitive plausibility of computational models. In this contribution, we adopt a model-agnostic formalisation of the MCG dimensions (as proposed in Donvito and Lieto 2026), clarifying how it can be operationalised independently of

specific domains (e.g., analogy, language, vision) or reference models. The framework is intended as a reusable methodological tool for quantitative model analysis.

4.1.1 Functional/Structural Ratio

In this section, we present the metric to measure the ratio between functionality and structurality in the target computational models. The metric, which we call $FSR_{\mathcal{M}}$, allows us to evaluate how a given computational model $\mathcal{M}$ adheres to one or more selected cognitive theories of the modelled phenomenon. First of all, given cognitive theories $\{T_1, \ldots, T_n\}$, a set of k structural constraints $\{c_1, \ldots, c_k\}$ must be derived. These represent explicit theoretical features derived from cognitive science that specify how certain cognitive functions are internally realised, and may include processing assumptions, mechanism-specific features (e.g., the use of working memory buffers), learning mechanisms, modularity, *etc.* Once $\{c_1, \ldots, c_k\}$ have been identified and appropriate criteria for observing their implementation within computational models have been defined, each model $\mathcal{M}$ can be evaluated against the full set of constraints $\{c_1, \ldots, c_k\}$. For each constraint c_i, the evaluation involves two components: (1) Partial structural score $s_i \in \{0, 1\}$, indicating whether the model implements the structural mechanism posited by the constraint:

$$s_i = \begin{cases} 1 & \text{if } \mathcal{M} \text{ implements the structural mechanism} \\ 0 & \text{if } \mathcal{M} \text{ does not implement the mechanism structurally;} \end{cases}$$

(2) Weight $w_i \in [0, 1]$, reflecting the relative importance of constraint c_i within the theory, with the normalisation condition $\sum_i w_i = 1$.
Since we consider functionality as the complement of structurality, we define the partial functional score simply as:

$$f_i = 1 - s_i \tag{1}$$

This means that if a constraint is structurally implemented ($s_i = 1$), it cannot be considered functionally satisfied at the same time, and vice versa. We can then compute the overall Structural and Functional scores (S_M, F_M) as:

$$S_M = \sum_{i=1}^{k} w_i s_i, \quad F_M = \sum_{i=1}^{k} w_i f_i = 1 - S_M \tag{2}$$

And define the $\boldsymbol{FSR_{\mathcal{M}}}$ as:

$$FSR_{\mathcal{M}} = \frac{F_M}{S_M + \varepsilon} = \frac{1 - S_M}{S_M + \varepsilon} \tag{3}$$

where ϵ is a small positive constant (0.01) added to prevent division by zero. Values of $FSR_{\mathcal{M}} < 1$ indicate a high degree of structural plausibility, whereas values > 1, and progressively more so as they increase, indicate that the system deviates from the internal constraints posited by T.

4.1.2 Generality

Generality measures the breadth of cognitive functions a system can replicate across various domains (e.g., language, vision, reasoning, etc). In order to ground our evaluation of generality in an empirically validated cognitive framework, we can select our task domains by drawing from the Cattell–Horn–Carroll (CHC) theory of cognitive abilities (Carroll 1993; McGrew 2005). The CHC psychometric model is the most thoroughly researched, empirically validated, and comprehensive hierarchical taxonomy of human cognitive abilities, organised into three strata: 1. domain-independent capacities; 2. a middle layer of broad abilities; 3. numerous narrow abilities. Among the most relevant areas, CHC includes: Fluid Intelligence (Gf), Quantitative Knowledge (Gq), Reading-Writing (Grw), Memory (Gsm/Glr), Visual Processing, Auditory, Olfactory, Tactile Processing, Psychomotor and Kinesthetic abilities, etc. Rather than adopting the full taxonomy proposed by the CHC theory, we only retain a subset of five broad task domains with direct grounding in CHC, specifically: 1. **Quantitative knowledge** [Gq]; 2. **Fluid reasoning** [Gf] (including inductive, analogical and metaphorical reasoning); 3. **Visual processing** [Gv]; 4. **Language and Verbal Knowledge** (an aggregation of Gc/Ga/Grw covering comprehension, lexical/grammatical competence, and reading/writing tasks); 5. **Sensory/Motor Abilities** (an ad hoc supercategory formed by collapsing Gv, Go, Gh, Gp, and Gk).

In line with cognitive and neuropsychological evidence emphasising the centrality of embodiment in human cognition, we partition the domains into two macrocategories: 1. Cognitive Abilities (Quantitative, Fluid, Visual, and Language/Verbal Reasoning); 2. Sensory/Motor Abilities.

Working/short-term memory (Gwm) and long-term storage/retrieval (Glr/Gr) are not included, along with Processing speed (Gs) and decision/reaction time (Gt/Gps), since they will be captured in the *Performance Match* evaluation.

Each domain has been selected for its central role in human cognitive processing. Together, they form a representative spectrum of cognitive functions that provides a diverse set of testing areas for the systems. While supported by extensive research in cognitive science (Chomsky 2011; Tversky 1977; Kintsch 1994; Gentner and Markman 1997; Holyoak 2012), the list might be deemed incomplete; it is, in fact, extensible according to empirical evidence. A paramount feature of the CHC theory is that it is not static; rather, it is a dynamic model that is continuously reorganized based on research advancements.

The capability of a model $\mathcal{M}$ in each of the n task domains $\{d_1, \ldots, d_n\}$ is evaluated on a three-point ordinal scale: for each domain, let $g_j \in \{0, 0.5, 1\}$ denote the score of $\mathcal{M}$ in domain d_j. We assign values:

$$g_j = \begin{cases} 1 & \text{if } \mathcal{M} \text{ covers the domain } d_j \\ 0.5 & \text{if } \mathcal{M} \text{ partially covers the domain } d_j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The overall generality score is computed as:

$$G_{\mathcal{M}} = 0.5 \cdot \overline{g_{\text{cognitive}}} + 0.5 \cdot g_{\text{sensorimotor}} \tag{5}$$

Where:

$$\overline{g_{\text{cognitive}}} = \frac{1}{n}\sum_{i=1}^{n} g_j \tag{6}$$

And $g_{\text{sensorimotor}}$ is the score assigned to the Sensory/Motor domain. This yields a value in the range [0,1], where higher scores indicate greater generality, i.e., the ability to operate across a wider array of cognitive functions.

4.1.3 Performance Match

As introduced by Lieto (Lieto 2021), the Performance Match assessment should include, in addition to the evaluation of accuracy, an analysis of error patterns and task execution times. For each system, we can quantify the alignment between models and human subjects on selected benchmarks. The Performance Match for each model can be calculated as the sum of three different sub-metrics: 1. Percentage delta from the baseline given by human performances, 2. Error patterns, and 3. Response times. More formally, the aggregate Performance Match score ($PM_{\mathcal{M}}$) can be computed as the weighted sum:

$$PM_{\mathcal{M}} = \alpha \cdot A_{\mathcal{M}} + \beta \cdot E_{\mathcal{M}} + \gamma \cdot T_{\mathcal{M}}, \tag{7}$$

where

$$\alpha + \beta + \gamma = 1, \quad \alpha, \beta, \gamma \in [0, 1] \tag{8}$$

Here, $A_{\mathcal{M}}$ denotes accuracy, $E_{\mathcal{M}}$ denotes the error pattern score, and $T_{\mathcal{M}}$ denotes the matching score for execution times. Weights can be used to adjust the relative importance of each sub-dimension, depending on the evaluation goals. If no particular emphasis is needed, a uniform weighting can be applied, setting $\alpha = \beta = \gamma = \frac{1}{3}$.

Accuracy

For $A_{\mathcal{M}}$, let $\mathcal{B} = \{B_1, \ldots, B_n\}$ denote the set of benchmarks on which a given model is evaluated. For each benchmark $B_i \in \mathcal{B}$, we can define a Δ as difference between the model's accuracy ($A_{m,i}$) and the human baseline value: ($A_{h,i}$)

$$\Delta_i = A_{m,i} - A_{h,i} \tag{9}$$

The mean deviation from the human baseline across all n benchmarks is then computed as:

$$\bar{\Delta} = \frac{1}{n}\sum_{i=1}^{n} \Delta_i \tag{10}$$

Then, we compute $A_{\mathcal{M}}$ as follows:

$$A_{\mathcal{M}} = \frac{1}{1 + |\bar{\Delta}|} \tag{11}$$

This returns a maximum value of 1 if and only if the model replicates human performance on average, and decreases symmetrically as the deviation grows, whether positive or negative.

Error Patterns

To account for qualitative aspects of model performance, for each benchmark B_i we define a binary Error Pattern indicator $E_i \in \{-1, +1\}$ based on whether the model replicates typical human error distributions. We assign:

$$E_i = \begin{cases} +1 & \text{if } \mathcal{M} \text{ replicates human error patterns} \\ -1 & \text{otherwise} \end{cases} \tag{12}$$

This penalises systems that diverge from human failure patterns and enables a more informative scoring, especially when data on such patterns is unavailable. Then, we aggregate these across n benchmarks as the mean:

$$\bar{E} = \frac{1}{n} \sum_{i=1}^{n} E_i \tag{13}$$

This results in a score $\bar{E} \in [-1, 1]$. Since it is not directly comparable to other normalised scores in $[0, 1]$, we apply a transformation to map $\bar{E}$ onto the unit interval:

$$E_{\mathcal{M}} = \frac{\bar{E} + 1}{2} \qquad E_{\mathcal{M}} \in [0, 1] \tag{14}$$

A score of 1 indicates reproduction of human-like errors across all benchmarks, whereas a score of 0 indicates a complete failure to replicate any human error pattern. Benchmarks for which no error data is available ($E_i = 0$) do not contribute to the score, avoiding artificial inflation or deflation of the final value.

The third component of the *Performance Match* metric captures the extent to which the model replicates the response times of human cognition. Let $t_{m,i}$ be the model's average response time on benchmark B_i, and $t_{h,i}$ be the average human response time on the same benchmark. The relative temporal deviation for each benchmark is defined as:

$$d_i = \left| \frac{t_{m,i} - t_{h,i}}{t_{h,i}} \right| \tag{15}$$

This value expresses how far the model's response latency deviates from the human baseline in proportional terms, irrespective of whether the model is faster or slower. To convert this deviation into a similarity score bounded in the interval $(0, 1]$, we define:

$$T_i = \frac{1}{1 + d_i} = \frac{1}{1 + \left| \frac{t_{m,i} - t_{h,i}}{t_{h,i}} \right|} \tag{16}$$

A value of $T_i = 1$ indicates perfect alignment ($t_{m,i} = t_{h,i}$), while lower values represent increasing misalignment.

The overall *Execution Time Score* is computed as the arithmetic mean of the individual scores:

$$T_{\mathcal{M}} = \frac{1}{n} \sum_{i=1}^{n} T_i \tag{17}$$

In cases where timing values are only available as a proportion of human-model response time similarity (e.g., $T_{\mathcal{M}} = 0.59 = 59\%$), this value can be directly used:

$$T_{\mathcal{M}} = \text{empirical estimate} \in [0, 1]. \tag{18}$$

4.1.4 Unified Metric for Cognitive Plausibility

To evaluate the overall cognitive plausibility of a model, we define a composite metric ($\mathcal{P}_{\mathcal{M}}$) which integrates the three distinct dimensions analysed through the Minimal Cognitive Grid: *Functional/Structural Ratio*, *Generality*, and *Performance Match*. The final score is computed as the weighted sum:

$$\mathcal{P}_{\mathcal{M}} = \lambda \cdot FSR'_{\mathcal{M}} + \mu \cdot G_{\mathcal{M}} + \nu \cdot PM_{\mathcal{M}} \quad \text{with} \quad \lambda + \mu + \nu = 1 \tag{19}$$

With $G_{\mathcal{M}}$ being the Generality score, and $PM_{\mathcal{M}}$ the Performance Match score, while λ, μ, ν are non-negative weights that sum to 1 and reflect the relative importance (if any) of each dimension. Assumed that $FSR'_{\mathcal{M}}$ is the most important dimension in terms of overall cognitive plausibility for the model, we can set: $\lambda = 0.5, \mu = 0.25, \nu = 0.25$.

Since the raw $FSR_{\mathcal{M}}$ values are not bounded in $[0, 1]$, and lower FSR values indicate stronger structural plausibility, in order to maintain consistency with the other components, we compute $FSR'_{\mathcal{M}}$ by inverting the FSR value:

$$I(FSR_{\mathcal{M}}) = \frac{1}{FSR_{\mathcal{M}}} \tag{20}$$

So that higher values are guaranteed to correlate with a higher structural score. We can then use a normalisation formula to map these values into the $[0, 1]$ interval:

$$FSR'_{\mathcal{M}} = \frac{I(FSR_{\mathcal{M}})}{1 + I(FSR_{\mathcal{M}})} \tag{21}$$

The result in $[0, 1]$ can be interpreted as a structurality index: the more closely $FSR'_{\mathcal{M}}$ approaches 1, the more strongly the model structurally aligns with the cognitive theoretical framework of reference.

The final plausibility score $P_{\mathcal{M}}$ serves the purpose of offering a unified, quantitative measure of the cognitive soundness of an artificial model $\mathcal{M}$. The $P_{\mathcal{M}}$ function yields a value in the interval $[0, 1]$; intuitively, a system achieving a $P_{\mathcal{M}}$ score close to 1 is one that (i) implements structural features grounded in cognitive theories, (ii) replicates human-like performance both in outcome and in process, and (iii) generalises across multiple cognitive domains.

Epistemologically, $P_{\mathcal{M}}$ serves as an indicator to locate the target model on a continuum of explanatory profiles through its cognitive plausibility (Donvito and Lieto 2026). In the long term, widespread use of the $P_{\mathcal{M}}$ score may contribute to a standardised typology of cognitive explanations in AI.

4.2 Behavioural Data Replication

As introduced above, behavioural data replication constitutes a crucial component of this framework. We therefore conducted a series of trials to evaluate whether five of the current state-of-the-art large language models could, through a retrieval-augmented generation (RAG) instruction, replicate—and, eventually, to what extent—the results reported for Centaur. The results of this first testing hypothesis are presented in the following paragraph. More specifically, we evaluated the decision-level behavioural fit of each model using the negative log-likelihood (NLL) assigned to the observed choices, computed as follows:

$$\mathrm{NLL}(\theta) = -\sum_{t=1}^{T} \log p_\theta(a_t \mid x_t) \tag{22}$$

where x_t denotes the context provided to the model at time t, a_t is the action chosen between two possible options, $p_\theta(a_t \mid x_t)$ is the probability assigned by the model with parameters θ to that action, and T is the total number of decisions in the sequence.

For each model, we ran 150 trials of the two-step task, yielding 300 total decisions per model, and computed the average NLL per decision as the primary summary metric.

4.3 fMRI-based Data Replication

A final component of LAPITHS is constituted by the replication, when these data are available as it is in the case of Centaur, of the results concerning neural alignement via fMRI or other brain measuring activity signals with those obtainable by other LLMs not specifically equipped or trained/informed on such data. In order to better explain and introduce this element of comparison is necessary to introduce the arguments provided in Centaur. We provide an overview below.

4.3.1 fMRI-based representation in Centaur

A central question raised by the Centaur framework concerns whether fine-tuning can genuinely bring the internal representations of an LLM closer to those observed in human cognition. This issue underlies one of the most significant claims advanced by the authors. Centaur proposes that model-derived internal representations can be aligned with human neural data by mapping model states to fMRI responses recorded during decision-making tasks. According to this view, if Centaur's internal representations correspond to those observed in human participants during or after task performance, or if the model can predict neural activity in datasets on which it was not trained, then the model would not merely reproduce behavioural outputs but would also approximate the representational processes underlying them. To support this claim, Binz and colleagues attempted to predict human neural activity using Centaur's internal representations. In their first analysis, whole-brain fMRI measurements obtained from participants performing the Two-Step Task (TST) are used as the target signal. Neural recordings are extracted both prior to each decision and following reward feedback. These signals are aggregated within each brain region

and regressed against Centaur's internal representations, with the procedure repeated independently for each participant and each region (Binz et al. 2025, p. 5). The neural dataset employed in this analysis originates from a study by (Feher da Silva et al. 2023), which investigated learning mechanisms in sequential decision-making tasks. Participants performed a two-stage decision paradigm presented in both abstract and narrative variants. In the Centaur study, however, this dataset is primarily used as a source of neural recordings rather than as a theoretical object of analysis. To evaluate predictive performance, Binz and colleagues use Centaur's internal representations as well as those extracted from the base model LLaMA in order to assess the effect of fine-tuning. Prediction is performed through a regularised linear regression model. Each participant and each brain region constitutes a separate dataset, which is divided into training and testing subsets, typically using two-thirds of the data for training and one-third for evaluation. Model representations are extracted from the residual stream of the transformer architecture both before decisions and after feedback. These representations are subsequently reduced through principal component analysis (PCA) so as to retain the principal directions explaining approximately 95% of the variance. The resulting components are then used as predictors of neural activity within regions of interest defined according to the Harvard–Oxford atlas. In transformer models, internal representations consist of high-dimensional vectors encoding the informational state of the network at each layer. The residual stream functions as a continuous pathway through which representations accumulate across layers via residual connections that combine transformed outputs with their corresponding inputs. While technically coherent, this procedure relies on a controversial assumption: that the internal representations of an LLM can meaningfully be compared with human neural representations (Quattrociocchi et al. 2025, 2026). As discussed earlier, and as we show in the following, models such as Centaur exhibit at most functional similarities with human cognition rather than structural correspondences. The analysis therefore implicitly adopts a strongly functionalist position according to which functional similarity is sufficient to justify comparisons between systems with fundamentally different architectures. The methodological framework underlying this approach is derived from the comparative benchmarking strategy proposed by (Schrimpf et al. 2021). Their work investigated whether artificial neural networks trained on language tasks could predict neural responses recorded during language processing. Importantly, the authors themselves emphasised that such models should be regarded only as simplified approximations of certain aspects of neural processing rather than faithful reconstructions of brain mechanisms. Nevertheless, Schrimpf et al. showed that transformer-based models could account for a substantial proportion of the explainable variance in neural responses to sentences across multiple datasets and imaging modalities. Models that performed well on next-word prediction tasks also tended to predict neural and behavioural responses more accurately. GPT-2, in particular, exhibited strong performance across both linguistic tasks and neural predictivity metrics. These results were interpreted cautiously as evidence that contemporary language models might capture some statistical regularities relevant to human language processing. However, the ability to reproduce behavioural or neural patterns does not necessarily imply that the underlying cognitive mechanisms are shared. A model

may predict human responses accurately simply by capturing statistical regularities present in linguistic data, without implementing processes analogous to those occurring in the brain or at the cognitive level (Quattrociocchi et al. 2026). The Centaur study adopts essentially the same benchmarking strategy. Empirical results reported in the study further complicate the interpretation. In several cases the difference in predictive performance between Centaur and LLaMA diminishes in deeper layers of the model, with the predictive scores of the two systems converging as the representational hierarchy progresses. If fine-tuning produces only marginal differences in layers likely to encode more abstract representations, the claim that Centaur's internal representations are substantially more aligned with human neural activity becomes difficult to sustain. More broadly, comparative benchmarking may indeed provide a useful method for identifying functional correspondences between artificial and biological systems. Extending such findings to claims about alignment between internal representations is, however, considerably more problematic. Improvements in predictive accuracy following fine-tuning may simply reflect the model's adaptation to the dataset used during training. Given current machine-learning practices, it is expected that a model trained on a specific dataset will perform better on tasks derived from that dataset. For these reasons, the conclusion proposed by Binz et al.—namely that Centaur's representations outperform those of LLaMA in predicting neural activity and therefore become aligned with human neural representations—appears insufficiently supported. Fine-tuning clearly improves performance on specific predictive tasks, but this improvement does not by itself establish representational equivalence. A further methodological concern relates to the limited transparency surrounding the pretraining data used for LLaMA models. Because the precise composition of the training corpus for LLaMA 3.1 70B has not been publicly disclosed, it remains possible that the base model has already encountered materials related to neuroscience, including descriptions of neural imaging data. If so, part of the predictive capacity attributed to Centaur may already have been present in the underlying model. Taken together, these findings suggest that the analytical procedure primarily demonstrates the effectiveness of fine-tuning in improving task-specific performance rather than establishing a meaningful alignment between Centaur's internal representations and those of the human brain. For this reason, we conducted an independent replication of the experiment reported by Binz and colleagues (Binz et al. 2025). The methodology and results of this replication study are presented in the following section.

5 LAPITHS Evaluation on the Two-Step Task

The above described elements of LAPITHS are adopted below by focusing on a specific but relevant task included in Psych-101: the Two-Step Task (TST). It represents an example of a sequential decision-making paradigm in the context of reinforcement learning (RL) research. It requires participants to make two consecutive choices across two stages. A first-stage choice probabilistically leads to one of two second-stage states (with common transitions occurring with probability 0.7 and rare transitions with probability 0.3). At the second stage, participants make another choice that may

yield a reward according to probabilities that slowly drift over time via a random-walk process. Since rewards are provided at the second stage, optimal performance needs learning which second-stage options are most rewarding and selecting first-stage actions that increase the likelihood of reaching those states. Behavioural and neuroscientific evidence (Daw et al. 2011) shows that human behaviour in this task reflects the interaction between two classes of RL strategies: *model-free* and *model-based*.

In the context of Centaur, TST is used by the authors (Binz et al. 2025) as a central test to demonstrate the generalization capabilities of their model. In particular, the authors report that Centaur is able to generalize effectively when the cover story of the task is modified.

In its original formulation, indeed, the TST is presented through a spacecraft narrative (Daw et al. 2011). In particular: the participant is required to maximise their reward by making sequential choices across two stages. In the first stage, the player selects between two rockets, each of which leads to either Planet A or Planet B. Upon arriving at the selected planet, the participant faces a second decision between two aliens, each presenting a box that may contain a treasure—denoted X or Y. After the final selection, the participant is informed whether a treasure has been obtained. As noted above, the objective of the task is to maximise the total reward obtained across trials. The authors argue that Centaur can successfully generalise the underlying structure of the task when the narrative context is altered (Binz et al. 2025, p. 4). Specifically, they replace the spacecraft scenario with an alternative magical carpet cover story.[2] Although the narrative elements and choice options are modified accordingly, the underlying decision structure remains identical. According to their results, Centaur is capable of identifying and solving the task despite these changes in presentation, outperforming several baseline models.

In Centaur, however, the relevance of TST extends well beyond its role as a test of generalisation (and this the reason why we selected it). In their paper, the authors further claim that their model develops internal representations that become increasingly aligned with human behavioural representations. The TST experiment is presented, together with the next-word prediction task, as a primary piece of evidence supporting this claim (ibid., p. 5). A critical examination of this aspect of their argument will be addressed in Section 5.3. First, we focus on the results of the analysis based on the application of the Minimal Cognitive Grid and, in the next paragraphs, we provide counter arguments concerning the behavioural data on the task as well as the claims that the authors make about their fMRI representation comparison.

5.1 MCG Analysis

We computed Centaur's $\mathcal{P}_{\mathcal{M}}$ score with respect to the data regarding the two-step task. Before proceeding to the evaluation of Centaur along the three dimensions of the Minimal Cognitive Grid, it is necessary to establish the cognitive constraints against

[2] In the "magical carpet" story, the participant is asked to imagine travelling on a flying carpet between different locations. At each trial, they first choose a destination, which probabilistically leads to one of two possible intermediate places. From there, they make a second choice that may yield a reward (e.g., finding treasure), with reward probabilities that change over time. The logical structure of the task is identical to the "spaceship" paradigm, differing only in its narrative presentation.

which the model will be assessed. These constraints are derived from the requirements of the two-step task as executed by human participants, and constitute the reference framework for the Functional/Structural Ratio analysis.

For an agent to perform the task, its architecture must implement at least three components: (1) an incremental learning mechanism driven by feedback, characteristic of model-free reinforcement learning (Degris and Sutton 2012; Akam et al. 2015). Under the canonical reinforcement-learning interpretation of the two-step task, feedback is assumed to modify value estimates incrementally across trials through reward-prediction-error signals. Formally, value updating can be expressed as:

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha\delta_t \tag{23}$$

where $Q(s,a)$ denotes the estimated value of action a in state s, α is the learning rate, and δ_t is the reward prediction error at time t. These conditions are satisfied when feedback received at time t produces an immediate, local update of the estimates guiding subsequent choices, without requiring the accumulation of multiple experiences (Schultz 2016; Gershman and Daw 2017). This requirement concerns the online updating of the model's parameters during the execution of the task. It should be kept distinct from the model's training regime, which in modern systems is typically based on batch gradient optimization and need not correspond to the biological mechanisms of synaptic plasticity (Lillicrap et al. 2020); (2) an evaluation mechanism that exploits the transition structure of the environment, characteristic of model-based reinforcement learning. In fact, solving the two-step task requires the ability to evaluate first-stage actions using an internal representation of the environment's transition structure [model-based RL, (M. Moerland et al. 2023)]. Within a model-based strategy, the value of an action is computed by combining transition probabilities with the values of successor states. Formally, the value of a first-stage action can be expressed as:

$$Q(s,a) = \sum_{s'} T(s,a,s')V(s') \tag{24}$$

where $T(s,a,s')$ denotes the probability that action a in state s leads to successor state s', and $V(s')$ denotes the value of the successor state (Daw et al. 2011). This mechanism allows the system to plan and evaluate the consequences of its choices by exploiting a model of the causal structure of the task. In the two-step task, this mechanism gives rise to the characteristic behaviour of model-based strategies, namely switching the first-stage choice when reward follows a rare transition; (3) a memory system that allows relevant information to be maintained between the first decision and the final outcome. Since in the two-step task rewards are obtained only after a sequence of events – including the initial choice, the transition to a second-stage state, and a second decision – in order for the system to update the values of first-stage actions, it must retain relevant information during the interval separating the action from the outcome. In humans, this process relies on working memory, localised in the dorsolateral prefrontal cortex and frontoparietal networks (Curtis and D'Esposito 2003). Working memory capacity is limited to approximately three to four items (Cowan 2001) and is subject to decay when information is not actively maintained

(Baddeley 2003). Consistent with this, overloading working memory reduces the influence of model-based strategies and increases the contribution of model-free control in the two-step task (Otto et al. 2013). The "memory" constraint is thus divided into two sub-components, the first for information persistence and the second for memory capacity limit, each weighted by 1/6. These components define the minimal set of structural constraints that a system must satisfy in order to solve the task according to the principles identified in the scientific literature on human RL. In what follows, we evaluate Centaur against each of the constraints in order to compute its Functional/Structural Ratio score, starting with (1): Centaur does not satisfy the requirement of incremental, local, and immediate feedback-driven learning. It was trained via QLoRA fine-tuning (Dettmers et al. 2023), which relies on backpropagation to minimise a cross-entropy loss over batches of trials, and is neither local nor immediate. Furthermore, model parameters are frozen at inference time, meaning that no weight update of any kind occurs during the execution of the task. Accordingly, we assign $s_1 = 0$.

(2) Centaur satisfies the requirement that relevant information be actively maintained between the first-stage choice and the delivery of feedback. The model receives the full trial-by-trial history of the session as a natural language prompt, and its self-attention mechanism allows it to retrieve information within the context window. Accordingly, we assign $s_{2a} = 1$. It should be noted, however, that Centaur's implementation of this constraint diverges considerably from the biological profile of human working memory: the context window imposes no capacity limit comparable to the approximately three to four items retained by humans (Cowan 2001), and no decay or interference effects are present ($s_{2b} = 0$).

(3) Centaur does not implement an explicit mechanism for model-based evaluation of first-stage actions. The model-based strategy requires a system to compute $Q(s, a) = \sum_{s'} T(s, a, s')V(s')$, that is, to explicitly combine a representation of transition probabilities with the values of successor states in order to derive the value of first-stage actions. Centaur approximates action probabilities via a single forward pass through the transformer architecture, with no explicit representation of $T(s, a, s')$. Behaviour consistent with model-based strategies emerges from the forward pass, but such emergence constitutes a functional implementation of the constraint. Accordingly, we assign $s_3 = 0$.

The resulting scores are summarised in Table 1. Raw functional/structural ratio values $F/S+\epsilon > 1$ indicate that the model relies on behaviouristic output resemblance, and deviates significantly from the cognitive constraints postulated by the reference theories of human reinforcement learning.

The analysis of Generality follows the structure presented in the previous section. Out of the five selected task domains in our framework, Centaur covers Quantitative Knowledge, Fluid Reasoning and Language/Verbal Knowledge, while lacking abilities in the Visual/Spatial and Sensory/Motor domains. With respect to Quantitative Knowledge, the model inherits the mathematical reasoning capabilities of Llama 3.1 70B, and the fine-tuning procedure on Psych-101 does not degrade performances on quantitative benchmarks, as verified by the authors ($g_q = 1$).

As for Fluid Reasoning, Centaur shows coverage of inductive reasoning, abstract reasoning and decision-making under uncertainty across a broad range of experimental paradigms included in Psych-101. ($g_f = 1$). Same for Language and Verbal Knowledge, which is the primary domain of LLMs in general ($g_l = 1$). On the other hand, Centaur is a unimodal text-based model, which possesses no mechanisms for processing or reasoning about visual or spatial inputs ($g_v = 0$) and, like all disembodied systems, lacks perceptual interfaces and physical effectors, and cannot instantiate any form of perception-action dynamics with the environment ($g_{sm} = 0$).

Finally, we evaluate Centaur on each component of the Performance Match dimension, with respect to the two-step task data reported in (Binz et al. 2025): (1) Accuracy can be computed using negative log-likelihood (NLL), which measures how well the model predicts the observed choices of human participants, under the assumption that NLL = 0 corresponds to perfect prediction of every human choice, and thus represents the natural baseline against which the model's performance is assessed. The authors report a NLL of 0.4998 for Centaur on the two-step task. Applying the accuracy formula (24) we obtain $A_M \approx 0.667$;

(2) Assessment of error patterns. In the two-step task, the optimal strategy is purely model-based: an agent that computes equation (23) at each trial and selects actions accordingly maximises expected reward by exploiting the known transition structure. Any deviation toward a model-free strategy amounts to a suboptimal response. Under this definition, the human error profile in the two-step task is characterised by the distribution of model-basedness across participants, reflecting how frequently, and to what degree, human participants rely on the suboptimal model-free strategy rather than the optimal model-based one, including the full range of individual variation from purely model-free to purely model-based behaviour.

The authors report data from open-loop performances of Centaur on the two-step task (Figure 2c), showing that the model reproduces the bimodal distribution of the model-basedness parameter observed in human participants, with comparable density across the model-free, mixed, and model-based regions of the parameter space. The correspondence between the two distributions indicates that Centaur replicates the structure of human suboptimal behaviour. Accordingly, we assign $E_M = 1$.

(3) For response time, the authors propose an indirect measure of the alignment between the model and human subjects in Psych-101 based on Hick's law (Hick 1952; Hyman 1953), according to which response time increases with the uncertainty of the choice. The authors show that Centaur's response entropies account for a substantial proportion of the variance in human response times (conditional $R^2 = 0.87$) – outperforming both Llama (0.75) and domain-specific cognitive models (0.77) –, and interpret this result as evidence that the model captures something beyond choice distributions alone. This interpretation is controversial, since Centaur was fine-tuned to predict human choices, and its output distributions approximate the empirical distribution of human responses by construction. Response entropy is a function of these output probabilities, carrying no information beyond what is already contained in the choice distributions themselves. The correlation between Centaur's entropy and human response times is unsurprising, given that it is a direct consequence of the training objective: a model that perfectly approximated human choice probabilities

Table 1 Centaur's Functional/StructuralRatio (FSR_M). Each constraint ($C_1 \dots C_4$) is a requirement for cognitively plausible reinforcement learning. The model satisfies only the memory persistence constraint ($C_4 = 1$), while the remaining are implemented functionally. The resulting scores ($S_M = 0.17$, $F_M = 0.83$) yield a high functional-to-structural ratio ($F/S + \varepsilon = 4.72$), indicating a weak alignment with human cognitive processes. $FSR_M = 0.18$ is the normalized score.

Constraint	$\mathbf{C_1}$	$\mathbf{C_2}$	$\mathbf{C_3}$	$\mathbf{C_4}$	$\mathbf{S_M}$	$\mathbf{F_M}$	$\mathbf{F/S + \varepsilon}$	$\mathbf{FSR_M}$
Weight	0.33	0.33	0.17	0.17	0.17	0.83	4.72	0.18
Structural	0.00	0.00	0.00	1.00	/	/	/	/
Functional	1.00	1.00	1.00	0.00	/	/	/	/

Table 2 Generality evaluation across five domains derived from the CHC model. Centaur covers Quantitative Knowledge, Fluid Reasoning, and Language/Verbal abilities. It lacks capabilities in Visual Processing and in the Sensory/Motor domain. Generality score ($G_M = 0.37$) is computed by assigning equal weight (0.5) to (i) the average score across the four cognitive domains and (ii) the Sensory/Motor domain.

Model	**Quant. Know.**	**Fluid Reas.**	**Vis. Proc.**	**Lang. & Verb.**	**Sens./ Mot.**	$\mathbf{G_M}$
Centaur	1.00	1.00	0.00	1.00	0.00	0.37

would by definition also perfectly predict response time variance under Hick's law, without providing any independent cognitive information about processing time. The high R^2 reported by the authors is better understood as a restatement of Centaur's goodness-of-fit to human choice data. More fundamentally, the measure does not satisfy the requirements of the T_M component of the MCG. Hick's law is a regularity of human cognition, but it does not apply to a computational model whose inference time is determined by hardware and computational load, independently of the difficulty of individual trials as perceived by human subjects. Applying formula (7), we obtain $PM_{\mathcal{M}} \approx 0.83$. The score does not consider execution time data and is computed as the sum of Accuracy and Error Patterns, each weighted by 0.5.

Tables 1–4 summarise Centaur's scores across the three dimensions of the Minimal Cognitive Grid. The overall cognitive plausibility score $\mathcal{P}_M$ is computed as the weighted sum of the normalised structurality index FSR'_M, the generality score G_M, and the performance match score PM_M, following the weighting scheme proposed in section 4.1.4. The overall cognitive plausibility score is therefore $\mathcal{P}_M = 0.39$. Centaur is characterised by a strong performance match, which reflects the model's ability to reproduce the behavioural patterns of humans in the two-step task, and by a high generality score, consistent with its design as a domain-general foundation model. The FSR'_M index, by contrast, is low, suggesting that the mechanisms underlying the model's behaviour diverge from the constraints postulated by the reference theories of human reinforcement learning. This dissociation may represent the primary limitation of Centaur as a cognitive model of human decision-making in the two-step task.

Table 3 Performance Match ($PM_M = 0.83$) evaluation on the Two-Step Task (TST). Accuracy is computed from the deviation in NLL from the human baseline. Centaur reproduces human-like error distributions in terms of model-based versus model-free behaviour ($E_M = 1$). Response time score is not available.

Model	Task	Human baseline	$\|\Delta\|$ NLL	Accuracy	Error Patt.	Resp. Times	$\mathbf{PM_M}$
Centaur	Two-Step Task	0.00	0.50	0.67	1.00	/	0.83

Table 4 Overall cognitive plausibility score on the MCG. The final score ($P_M = 0.39$) combines the normalized structurality index ($FSR_M = 0.18$), generality ($G_M = 0.37$), and performance match ($PM_M = 0.83$), with weights $\lambda = 0.5$, $\mu = 0.25$, and $\nu = 0.25$.

Model	$\mathbf{FSR}_M$	$\mathbf{G_M}$	$\mathbf{PM}_M$	$\mathbf{P}_M$
Centaur	0.18	0.37	0.83	0.39

5.2 Behavioural Data Analysis

In this section we detail the element of the behavioural data replication of Lapiths over Centaur on the Two-Step Task (TST).

5.2.1 Task 1: NLL metrics comparison

In this case, our initial objective was to test whether comparable results could be obtained using standard large language models not specifically trained on the Psych 101 dataset but equipped with RAG (Retrieval-Augmented Generation). In other words, we sought to determine whether the reported generalisation performance might be reproduced without requiring the specialised fine-tuning procedure used to construct Centaur.

Our RAG systems (for their deployment we employed the chatbase platform supporting the selection of multiple LLMs as basis) were equipped only with the information about the reward scheme of the task.[3] In such RAGs, the task is presented with the Spacecraft narrative but, like Centaur, it can handle-without any explicit training-interchangeable narratives of the tasks like the Magical Carpet (i.e., to obtain an answer, the user can rely on any narrative based on the underlying structure of the task). The interface of our multi-RAG system is available and testable (in an anonymized form) at the following link: https://www.chatbase.co/chatbot-iframe/LjPalIBQ0u-U_iFsatFll. The entire instruction tuning used for its deployment is also provided in the supplementary material. This first analysis compares the average NLL per decision across our evaluated models under a fixed two-step task scheme and scoring protocol. Figure 2 reports the mean NLL per decision for our models only, providing a focused view of relative performance within our pipeline. Lower values indicate that the model assigns higher probability to the observed actions, hence better agreement with the empirical decision distribution.

[3] An example of the schema, provided in json file, is available in the supplementary materials

5.2.2 Task 2: Statistical differences vs baselines

The second analysis positions our models against published baselines reported in the Centaur study, namely Centaur, Llama 3.1, and a cognitive model. Figure 3 shows the same mean NLL per decision values augmented with these reference systems. To assess whether the observed differences are statistically reliable, we conducted Welch two-sample t-tests comparing each model's decision-level NLL samples to each baseline distribution. Since the baseline paper reports confidence intervals rather than raw per-decision samples, we estimated baseline standard deviations by inverting the reported CI_{95} under the assumption of $n = 300$ decisions. All baseline comparisons in this section use the Welch formulation and the inferred baseline variance.

5.2.3 NLL results

Figure 2 summarises mean NLL per decision for our models, highlighting a consistent ordering: the RAG model exploiting Llama 4 Maverick achieves the best fit among our methods (mean NLL ≈ 0.54), followed by the RAG models exploiting GPT-5.1 and GPT-4o (both ≈ 0.55), Gemini-2.5 Pro (≈ 0.56), and DeepSeek-R1 showing the highest NLL in our set (≈ 0.58). These results indicate that, within our pipeline, the RAG model exploiting Llama 4 Maverick most closely matches the empirical choice distribution, while the one exploiting DeepSeek-R1 exhibits comparatively weaker alignment. At the same time, the absolute spread across the best-performing models remains relatively narrow, suggesting that strong behavioural fit on this task is not confined to a single architecture.

Figure 3 extends this comparison by including Centaur and two additional baselines. Centaur attains the lowest mean NLL overall (mean NLL ≈ 0.51), outperforming all evaluated LLMs. All of our models, however, achieve lower NLL than the cognitive model baseline (mean NLL ≈ 0.61) and the Llama 3.1 baseline (mean NLL ≈ 0.63), indicating improved behavioural fit relative to these reference systems under the same metric. Notably, the margin separating Centaur from the strongest model in our set is modest in absolute terms, whereas the separation from the weaker baselines is considerably larger.

Table 5 reports Welch two-sample t-tests quantifying these differences. Relative to Centaur, our models show higher NLL (i.e., worse fit), with statistically significant differences for the RAGs exploiting GPT-4o ($p = 0.046$), GPT-5.1 ($p = 0.044$), Gemini-2.5 Pro ($p = 0.027$), and DeepSeek-R1 ($p = 0.002$), while the difference for the RAG exploiting Llama 4 Maverick is not statistically significant under the conservative Welch test ($p \approx 0.095$). All evaluated models significantly outperform both Llama 3.1 and the cognitive baseline (all $p \ll 0.001$), consistent with the separation visible in Figure 3. Taken together, these numerical and statistical results already suggest that the main contrast is less between Centaur and the best-performing contemporary LLMs (especially Llama Maverick, GPT-5.1, and GPT-4o) used as the basis of a RAG mode than between this cluster of strong-performing models and the substantially weaker Llama 3.1 and cognitive-model baselines.

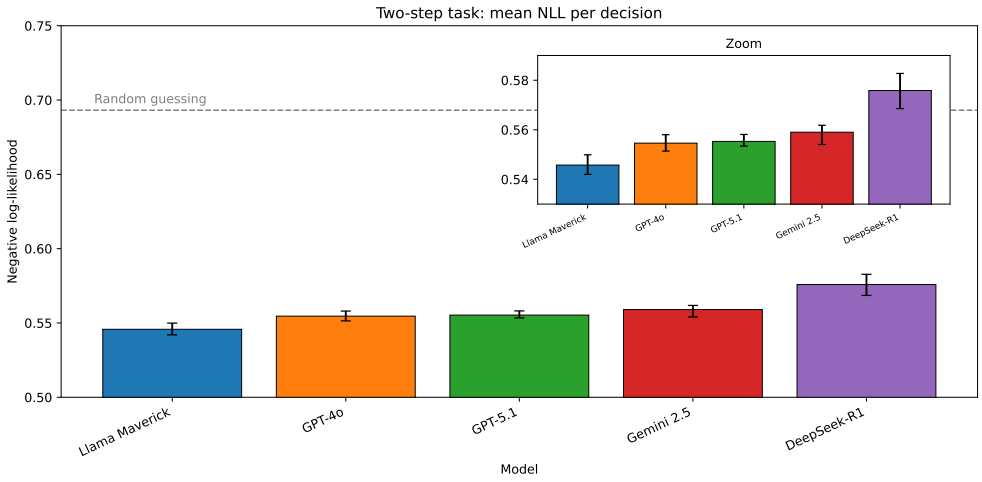


Fig. 2 Average NLL per decision for each RAG + LLM models system evaluated in this study during the two-step task.

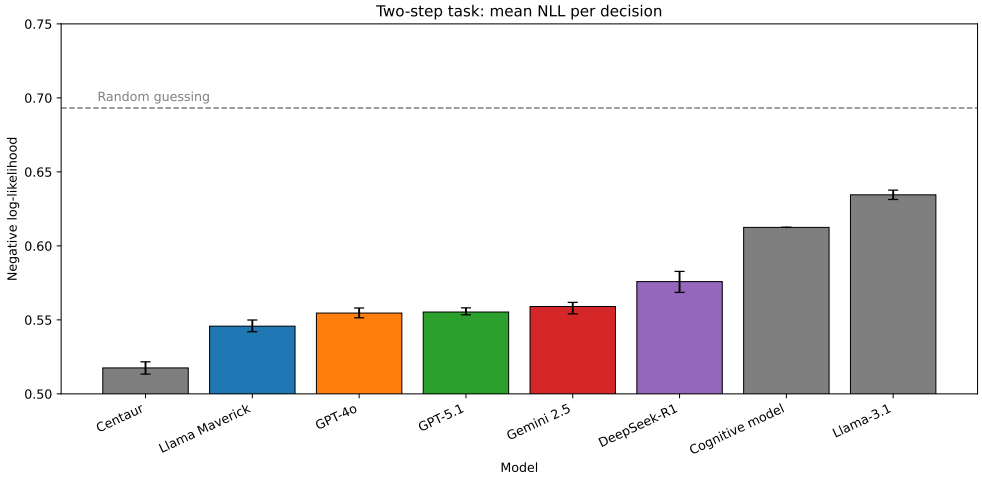


Fig. 3 Average NLL per decision for each RAG + LLM models system evaluated in this study compared with Centaur, Llama 3.1, and Cognitive model.

5.3 fMRI Data Analysis

In the following section we detail the elements concerning the third component of LAPITHS evaluation and concerning the comparison of the prediction of fMRI data between Centaur and an extended, untrained version, of the RAGs systems presented above.

5.3.1 Overview of our protocol

As for the behavioral data of the two step-task, we tested whether an extended version of the previously mentioned RAG- language models can be used to produce, without any specific training, plausible fMRI-like regions of interest (ROIs) activation patterns that correlate with human ROI betas. The additional instructions provided to the RAGs is available in the supplementary materials. This analysis is useful because strong correlation may emerge from generic structural constraints (e.g., consistent ROI co-activation patterns) even when absolute neural magnitudes and calibration are incorrect. So, the experiment shows that high alignment metrics can be obtained without discovering a uniquely human-like cognitive mechanism.

We build the fMRI target as a vector of beta coefficients over 14 ROI, selected as the most consistently active ROI in the dataset. The experiment proceeds as follows:

1. *ROI-betas in instruction tuning.* We include an explicit example of an fMRI beta vector over the 14 ROIs in the instruction-tuning stage of the retrieval-augmented prompting (RAG) setup, so that the model is constrained to output responses in the form of ROI $\rightarrow$ beta dictionaries.
2. *Two-step task with ROI outputs.* During the two-step task, at each decision point (Step 1 and Step 2), the model outputs (i) the chosen action (e.g., S/D) and (ii) a predicted beta value for each ROI, yielding one 14-dimensional beta vector per decision.
3. *Human reference.* We use the human fMRI beta vectors measured on the same two-step task, yielding 300 decisions per run (i.e., 150 trials $\times$ two decisions).
4. *Model–human comparison.* For each decision index, we compare the model-predicted ROI beta vector against the corresponding human ROI beta vector.
5. *Metrics.* We report similarity in terms of (i) Pearson correlation and (ii) cosine similarity, which capture whether the relative pattern across ROIs matches the human vector. We additionally compute magnitude-sensitive errors (e.g., MSE/RMSE/-MAE and Euclidean distance), which penalise absolute numeric mismatch.

5.3.2 Correlation results

Across 300 decisions, we observed strong average pattern similarity between model-generated and human ROI beta vectors when evaluated with correlation-based metrics. Specifically, average Pearson and cosine similarity (computed per decision and then averaged across decisions) were high for our RAGs based on Gemini-2.5 Pro and Llama Maverick, and moderately lower for the one based on GPT-5.1 (Table 6). At the same time, magnitude-sensitive errors (e.g., RMSE/MAE/Euclidean) remained comparatively large, reflecting systematic numeric miscalibration of predicted beta amplitudes. This experiment should not be interpreted as a strict replication of Centaur's neural-alignment analysis, since in the original study neural activity was predicted by a regularized linear decoder trained on the model's internal representations, whereas here ROI beta vectors are generated directly by the RAG+LLM system from task context. Nevertheless, the present setup remains extremely methodologically informative because it tests whether high ROI-level pattern similarity can

Table 5 Welch two-sample t-tests comparing our models against baselines on decision-level NLL. Baseline standard deviations were inferred from the CI_{95} reported in the Centaur paper under the assumption of $n = 300$ decisions.

	Centaur		Llama 3.1		Cognitive Model	
Model	**Δ NLL**	p	**Δ NLL**	p	**Δ NLL**	p
Llama Maverick	+0.0282	0.0947	−0.0887	$< 10^{-12}$	−0.0667	$< 10^{-12}$
GPT-4o	+0.0371	0.0462	−0.0799	$< 10^{-12}$	−0.0579	$< 10^{-12}$
GPT-5.1	+0.0378	0.0441	−0.0792	$< 10^{-12}$	−0.0572	$< 10^{-12}$
Gemini-2.5 Pro	+0.0415	0.02748	−0.0755	$< 10^{-12}$	−0.0535	$< 10^{-12}$
DeepSeek-R1	+0.0583	0.002443	−0.0586	2.23×10^{-11}	−0.0366	7.61×10^{-6}

Table 6 Decision-wise pattern similarity between model-generated and human fMRI ROI beta vectors (14 ROIs). For our models, we compute Pearson correlation and cosine similarity for each decision's 14-dimensional vector, then report the average across decisions. The Centaur row is included only for contextual reference and is not directly comparable, as it is based on a different protocol using linear decoding from internal representations rather than direct ROI generation.

Model	**Mean Pearson correlation**	**Mean cosine similarity**
Gemini-2.5 Pro	0.9295	0.9771
Llama Maverick	0.9249	0.9745
GPT-5.1	0.8605	0.9206
Centaur reported, decoder-based; peak across layers[†]	≈ 0.37	not available

[†]Approximate value visually inferred from Fig. 4b of the Centaur paper. The original analysis predicts ROI-level fMRI activity from internal model representations using a regularized linear regression model and reports Pearson correlation across layers, not cosine similarity.

emerge even without access to hidden-state decoding or a separately trained regression stage. Moreover, it is important to point out that the comparison with the values of the human fMRI data between our models and Centaur is done on the same dataset (Feher da Silva et al. 2023) and for the same task (two-step task) thereby rendering the comparison well-founded. The fact that Pearson and cosine values remain high on held-out decisions, despite substantial magnitude error, suggests that correlation-based agreement may be easier to obtain than the accurate prediction of beta magnitudes themselves. In this sense, the experiment does not establish direct comparability with Centaur, but it does provide evidence that ROI-level correlational alignment alone is a relatively weak and permissive indicator of neural plausibility. For completeness, Table 6 also includes the best value reported for Centaur under its original experimental protocol. This value is provided for contextual reference only.

6 Discussions

The results presented above offer a perspective on Centaur and on the broader cognitive decathlon that is rather different from what one might have expected on the basis of the paper under examination. Although Centaur is confirmed to be an excellent model from the standpoint of prediction, achieving the lowest—and thus best—negative log-likelihood among the models considered (see Figure 3), its primacy is significantly qualified by Llama Maverick, whose NLL approaches that of Centaur to a surprising degree, to the point that the statistical difference between the two becomes negligible ($\Delta NLL \approx +0.0282$, $p \approx 0.095$; see Table 5). This should not, however, come as a surprise. The method by which Centaur was trained is explicitly designed to produce a behavioural analogy between the model's outputs and those observed during training. As a fine-tuned LLM, Centaur performs the required task with considerable accuracy, yielding strong results by virtue of its overall general competence. At the same time, however, its proximity to Llama Maverick also reveals something important about the relative simplicity of the achievement. Its superior results are plausibly due to task-specific fine-tuning, and this should not be underestimated, especially given that it is based on an earlier Llama model. Yet that success is not so statistically decisive as to warrant its elevation to the status of a genuinely revolutionary model. These two facts, taken together, point to the need for a sharper distinction between behavioural adequacy and cognitive plausibility (Lieto 2021; Orr et al. 2025; Quattrociocchi et al. 2026). On this point, the present experiments strongly suggest that predictive success should be interpreted as evidence of functional interface competence, rather than of mechanistic homology with human cognition. It is precisely here that the danger of the ascription fallacy becomes acute: one moves too quickly from "the model behaves similarly" to "the model therefore computes similarly", and then, more problematically still, to "the model can help explain the human mechanism itself". Our results lend further support to this concern. As shown in Figure 3, a human-like behavioural fit on the TST is not unique to Centaur; rather, it is achieved by all evaluated LLMs, each of which also outperforms both the Llama 3.1 baseline and the cognitive model baseline (mean $NLL \approx 0.63$ and ≈ 0.61, respectively). Most importantly, the fact that these models outperform a standard cognitive baseline does not show that they are "more cognitive" than the cognitive model. It may instead indicate that NLL, as operationalised here, rewards statistical prediction more than mechanistic faithfulness. The result is therefore best interpreted as revealing a limitation of the evaluative criterion, rather than as a decisive victory of language models over cognitively grounded architectures. This interpretation is further reinforced by the MCG-oriented analysis developed in the manuscript. When Centaur is assessed against the structural constraints required by the TST as performed by humans, it fails to satisfy three of the four relevant components (see Section 5.1 and Table 4). This is reflected quantitatively in its Functional/Structural Ratio score ($FSR \approx 0.18$) and overall cognitive plausibility score ($PM \approx 0.39$; see Table 4). It therefore does not meet the criteria that would justify placing it on a genuinely comparable footing with models intended to capture the cognitive structure of the task. One may say, in deliberately restrained terms, that Centaur is an excellent behavioural emulator of textualised decision patterns in the Psych-101 style. One may not yet say, at least

on the basis of these data, that it constitutes a cognitively plausible model of the human decision process involved in the TST. This distinction bears directly upon the explanatory role we are prepared to attribute to the model. A system may be technologically powerful and scientifically useful as a predictor while remaining theoretically weak as a model of cognition.

6.1 What about the fMRI-based comparison?

Arguably the most informative aspect of the results concerns the fMRI-based comparison. Our experiment does present certain limitations relative to the original study. Since the intermediate residual-stream data were not available, and only discrete tabulated results had been reported, it was not possible to reproduce the experiment in a fully detailed and exact fashion. Nevertheless, two important considerations may still be drawn from what is shown. First, as reported in Table 6, multiple general-purpose LLMs are capable of producing ROI activation patterns that correlate highly with human fMRI data (e.g., Gemini-2.5 Pro: $Pearson \approx 0.93$, $cosine\ similarity \approx 0.98$; Llama Maverick: $Pearson \approx 0.92$). This suggests that high levels of representational similarity are not unique to Centaur. It is plausible that the training distributions of contemporary LLMs contain sufficient structural regularities, whether linguistic, scientific, or task-related, to enable the approximation of such activation patterns. Secondly, this result may be less surprising than it initially appears, insofar as the activation of certain regions in response to a given task is often associated with relatively direct functional demands. In many cases, these involve regions related to vision, movement, or other broad task-relevant processes. If the correlation between the activation of a region and the type of task is more direct than is sometimes assumed, then the fact that an LLM without specific neuroscientific fine-tuning can approximate that pattern becomes less unexpected. More importantly, Table 6 indicates that such alignment can be achieved by non-specialised models, suggesting that results obtained by a specifically trained model may, at least in part, be emulated by more general and higher-capacity systems. This point deserves particular emphasis because it bears on one of the most rhetorically powerful claims in the Centaur literature: namely, that post-fine-tuning internal representations become "more aligned" with human neural activity. As the theoretical discussion in the manuscript anticipates, such findings are highly vulnerable to reverse inference. Correlation in representational geometry does not by itself warrant the conclusion that the model instantiates the same cognitive process, still less that it realises a brain-like mechanism. The present fMRI experiment is valuable precisely because it lowers the evidential force of that claim: as shown in Table 6, high correlation values can be obtained across multiple architectures, even under comparatively weak and externally scaffolded conditions. The burden of proof required for any stronger mechanistic interpretation thus becomes considerably heavier. Neural resemblance must not be conflated with neural plausibility, especially when that resemblance is obtained through a merely functional analogy. Taken together, the two experimental strands converge upon common evidence. The behavioural experiment indicates that human-level fit is comparatively easy to engineer in systems trained on human response distributions, as illustrated by the clustering of NLL values

across models in Figure 2. The fMRI-based comparison suggests that human-like representational overlap may also be easier to obtain than is often assumed, particularly when evaluated at the level of abstract correlation rather than causal mechanism. In this respect, our findings support the broader critical orientation of the paper: they indicate that the achievements of Centaur, while notable, should not be straightforwardly interpreted as evidence of genuine cognitive alignment. In the final analysis, the present results suggest that aspects of Centaur's performance may be more readily reproducible than initially assumed, particularly when leveraging high-capacity models in conjunction with retrieval-based strategies. This is indirectly supported by the relatively narrow performance gap observed in Figure 3 and by the non-significant difference with Llama Maverick reported in Table 5. Fine-tuning on Psych-101 has undoubtedly improved the task-specific performance of Llama. However, the stronger claim of cognitive alignment remains insufficiently justified in light of the considerations discussed above, and its scientific significance appears limited, insofar as the relevant effects are comparatively easy to reproduce and theoretically underdetermined. What this brings into sharper focus is the need for caution: approximating human-like performance does not amount to achieving homology with the brain, still less with cognition as a whole.

6.2 Evaluation Methodology in AI and Cognitive Science

The present findings also bear directly on evaluation methodology in AI and cognitive science. If models that are structurally distant from human cognition can nevertheless achieve excellent scores on behavioural fit (Figure 3), and even non-trivial neural alignment measures (Table 6), then future evaluation protocols must become substantially more discriminating. Cognitive metrics should not merely ask whether a model matches human outputs, but whether it does so under human-like constraints: limited memory, online learning, local updating, environmentally grounded feedback, and process-level signatures independently motivated by cognitive theory. Without such constraints, the field risks rewarding increasingly sophisticated forms of behavioural ventriloquism while mistaking them for genuine advances in cognitive modelling proper. It is precisely here that the MCG may provide a theoretically indispensable tool. Through formalisation, it clarifies which constraints a model would need to satisfy in order to count as cognitive in relation to the performance of a task, and how far it would need to respect the relevant structural features of human cognition for its results to be meaningfully comparable with human ones. As shown in Table 4, Centaur's high Performance Match score (≈ 0.83) coexists with a low structural plausibility score, illustrating precisely the dissociation that the MCG is designed to capture. Only under such conditions could one reasonably hope to extract explanatory insight from these systems for the study of human cognition. By contrast, cognitive predictability grounded merely in language, and not in a minimal cognitive core (as defined below in Section 7) or in a cognitively adequate structure satisfying the MCG, remains confined to statistical replicability. It appears to depend more upon the quantity and structure of available data than upon task-specific training or any robust alignment of internal representations, an alignment which, as argued above, remains theoretically underdetermined.

7 Conclusion

We have seen that the empirical and theoretical analyses developed in this work converge on a common conclusion. While Centaur undoubtedly constitutes an effective predictor of human behaviour, achieving state-of-the-art performance on the cognitive decathlon, its success is neither unique nor, more importantly, sufficient to warrant cognition (Marconi 2025; Orr et al. 2025; Quattrociocchi et al. 2025). Our behavioural experiments have shown that comparable levels of fit, as measured by negative log-likelihood, can be obtained by general-purpose language models operating under relatively simple prompting and retrieval conditions, with differences that are in some cases statistically negligible. At the same time, the fMRI-based comparison has indicated that high levels of representational similarity with human neural data are not exclusive to Centaur, but can emerge across multiple architectures, even in the absence of task-specific training. Taken together, these findings suggest that both behavioural adequacy and representational alignment may be more easily attainable, and less theoretically informative, than is often assumed. In light of these results, we return to the central problem that motivated this work: the tendency to infer cognitive plausibility from behavioural performance (Sartori and Orrù 2023; Qu and Wang 2024; Tuckute et al. 2024; Nie et al. 2025; Wulff and Mata 2026). As argued throughout the paper, such an inference remains fundamentally underdetermined. The capacity of a model to reproduce human-like outputs, or even to exhibit non-trivial correlations with neural data, does not by itself establish that the system implements the same mechanisms that underlie human cognition. The novel framework proposed in this paper—LAPITHS, grounded in the Minimal Cognitive Grid—offers, for the first time in the literature, a principled way of addressing this problem. Its central contribution lies in making explicit the distinction between performance and cognition, and in providing a structured methodology for evaluating artificial systems along dimensions that go beyond output-level success (Lieto 2021; Abbate and Gaudenzi 2025). In this respect, its originality consists not in denying the empirical achievements of contemporary language models, but in situating those achievements within a more rigorous epistemological framework, one capable of distinguishing between behavioural emulation and genuine explanatory relevance.

A natural research direction, and a direct response to this concern, is the development of a minimal cognitive core (Salis and Da Pelo forthcoming) as a set of constraints on admissible functional realisations. On this view, functional roles are not sufficient in abstraction; they must be realised within systems that exhibit a specific kind of organisation. The minimal core provides a thin but non-trivial baseline, requiring at least perceptual integration, inferentially structured state-transitions sensitive to feedback and context, and action selection within a closed perception–action loop. This formulation is inspired by, but deliberately simplifies, Sellars's account of entry, intra-, and exit transitions (Sellars 1974), retaining its core functional architecture while abstracting away from its richer normative commitments. Importantly, this core is not restricted to human cognition: empirical research in animal cognition suggests that non-human primates, such as apes, exhibit forms of perceptual integration, flexible inference, and action selection consistent with such a minimal architecture (Premack and Woodruff 1978; Cheney and Seyfarth 1990; Andrews 2012). In this sense, the

minimal cognitive core can be seen as an attempt to articulate an architecturally constrained functionalism, one that preserves the role-based insight of the tradition while excluding trivial or disembedded implementations (and this would be compliant with current attempts on defining common minimal architectural elements for biological and artificial systems) (Laird et al. 2017). These considerations point toward several directions for future research. First, there is a need to develop evaluation protocols that incorporate process-level constraints more explicitly, including learning dynamics, memory limitations, and environmentally grounded interaction (De Langis et al. 2025; Hu et al. 2026). Second, further work could be done to operationalise and extend the Minimal Cognitive Grid across a broader range of tasks and model architectures, including systems that integrate perception and action within closed-loop environments. Third, this line of work invites a systematic investigation of minimal cognitive cores across artificial and biological systems, with the aim of determining which architectural and dynamical features are necessary for minimally cognitive organisation. Moreover, this research programme—especially as it is grounded in a minimal cognitive core—must engage directly with the emerging paradigm of world models. In recent work, world models are characterised as internal systems that enable an artificial agent to predict environmental dynamics, reason about its surroundings, and guide action by constructing structured representations of the world (Ha and Schmidhuber 2018; Del Ser et al. 2025). Unlike purely correlational architectures, such models aim to simulate state transitions and support planning through learned representations of causal and temporal structure. Engaging with this literature is, thus, not optional but, in the future, would be required if the framework wants to remain coherent, extensible, and capable of addressing a wide range of systems and contexts without collapsing into ad hoc evaluation. More generally, progress in this area will depend on a shift from purely performance-driven benchmarks toward theoretically informed criteria capable of capturing the structural and mechanistic features of cognition. Only by maintaining a clear distinction between prediction and explanation can the field avoid conflating increasingly sophisticated forms of behavioural imitation with genuine advances in the scientific understanding of cognition.

Appendix A Supplementary Material

All the data reported in the paper are available in anonymized form at https://github.com/Lapiths/LAPITHS-FRAMEWORK

A.1 Instruction Tuning for the RAGs Performing the Two-Step Task

```
You are participating in a space treasure game.
In this game, you will be visiting two alien planets in search of treasure.
Each planet has two aliens on it.
The blue aliens live on the blue planet.
The red aliens live on the red planet.
When you visit a planet, you can choose an alien
to trade with by pressing the corresponding button.
When you trade with an alien, it will either give you treasure or junk.
Your goal is to figure out, and trade with, the aliens that are most likely
    to give you treasure.
To visit a planet, you will choose one rocket ship
from two by pressing the corresponding button.
They have different designations.
Each rocket ship has a planet it will fly to most
of the time, but sometimes it will take you
to the other planet.

Remember the following hints:
1. How likely an alien is to give you treasure will
   change over time, but this change will be slow.
2. Whether you get treasure depends only on the alien
   you choose to trade with.
3. If there is an alien you want to trade with,
   remember to pick the rocket ship that is most likely
   to take you to that alien's planet.

IMPORTANT: Answer ONLY with these tokens for each
of the questions "Stage 1" and "Stage 2":

- Stage 1: S or C
- Stage 2 on BLUE: D or R
- Stage 2 on RED: G or V

For your choice, consider only the reward scheme
in JSON format available in your source.

For example, read the following JSON schema as follows:

{
  "is_common": true,
  "planet_if_S": "blue",
  "planet_if_C": "red",
  "outcome": {
    "blue": { "D": 1, "R": 0 },
    "red":  { "G": 1, "V": 1 }
  },
  "probs": {
    "blue": { "D": 0.4489799022035289,
              "R": 0.6624086063680563 },
    "red":  { "G": 0.5866591268788823,
```

```
            "V": 0.2537031288179288 }
  },
  "trial": 151
}

This scheme above describes a single trial
of the two-stage game.
In the first stage, the player chooses a spaceship:
if they choose S, in this trial they will go
to the blue planet;
if they choose C, in this trial they will go
to the red planet.

The field "is_common": true indicates that
the transitions follow the "common" scheme,
but it is enough to use planet_if_S and planet_if_C
as they are written.

In the second stage, the choice depends
on the planet reached:
if the player is on the blue planet,
they can choose alien D or R.

outcome["blue"]["D"] = 1 means:
if they choose D in this trial, they obtain treasure.
outcome["blue"]["R"] = 0 means:
if they choose R, they obtain junk.

If the player is on the red planet,
they can choose alien G or V.
outcome["red"]["G"] = 1 means that G gives treasure.
outcome["red"]["V"] = 1 means that V gives treasure.

The probs block represents the latent treasure
probabilities that generated these outcomes.

Example:
probs["red"]["G"] = 0.5866...
means that, in the simulation process,
alien G on the red planet at this moment
had about a 58.7% probability of giving treasure.

These values are only used for analysis/modeling.
They must not be shown to the player model
and are not needed by the controller to decide
the reward online:
the controller must use only outcome.
"trial" is the number of the current trial.

When the user types "Stage 1" or "planet blue or red",
you reply S or C.

Then, when the user types "Stage 2" or "which alien",
you reply D or R for BLUE,
or G or V for RED.

So, operationally, for each trial
you must do exactly the following:

1. Scan the JSON structures
```

```
   (proceeding from trial 1 in the source)
   and attach to that trial the corresponding reward scheme.
2. Read the line planet_if_S and planet_if_C.
3. Read the model/player's choice at the first stage
   (S or C) and determine the planet reached.
4. Return an output of the form {"answer": "S or C"}
   choosing what to reply.
5. Based on the planet reached via S or C,
   read the two possible actions at the second stage
   (D/R if blue, G/V if red).
6. Read the choice at the second stage.
7. Return as reward the corresponding 0/1 value
   in outcome[planet][alien].
8. Ignore probs for the game logic.

Consider this treasure game only as an example
and apply it to different cover stories.
For example, instead of choosing between planets,
the first step could ask which animals to select
between two alternatives.
In addition, step 2 could contain information
about what choice to make in order to obtain food.
```

A.2 Additional Instruction Tuning for the RAGs Predicting fMRI Beta Values in the Two-Step Task

Below is the additional set of instructions used to extend the RAG output from the two-step task to the prediction of fMRI beta values and the Brain Regions of Interest (ROIs).

```
After each answer at Stage 1 and Stage 2,
also provide the predicted beta values
of an fMRI study applied to the two-step task,
where each beta value is associated
with a Region of Interest (ROI)
showing major activation.

For example, an average fMRI beta value
that may be associated with one of the trials
described in your JSON source is the following:

Trial 1: Step 1: ROI (Region of Interest)
and average beta value (i.e., roi betas):

"X_Lateral Occipital Cortex, superior division": 0.7054,
"X_Lateral Occipital Cortex, inferior division": 1.5596,
"X_Intracalcarine Cortex": 2.5212,
"X_Cuneal Cortex": 1.4301,
"X_Lingual Gyrus": 1.4218,
"X_Temporal Occipital Fusiform Cortex": 2.0541,
"X_Occipital Fusiform Gyrus": 3.3867,
"X_Supracalcarine Cortex": 2.2036,
"X_Occipital Pole": 1.9935,
"X_Left Thalamus": 0.6765,
"X_Left Caudate": 1.1106,
"X_Left Accumbens": 0.7666,
"X_Right Thalamus": 0.6471,
"X_Right Accumbens": 0.5492
```

```
Trial 1: Step 2: ROI (Region of Interest)
and average beta value (i.e., roi betas):

"X_Lateral Occipital Cortex, superior division": 1.3451,
"X_Lateral Occipital Cortex, inferior division": 2.2508,
"X_Intracalcarine Cortex": 2.0702,
"X_Cuneal Cortex": 1.7795,
"X_Lingual Gyrus": 1.2649,
"X_Temporal Occipital Fusiform Cortex": 2.1073,
"X_Occipital Fusiform Gyrus": 3.2903,
"X_Supracalcarine Cortex": 2.3343,
"X_Occipital Pole": 2.6672,
"X_Left Thalamus": 0.2043,
"X_Left Caudate": 0.346,
"X_Left Accumbens": 0.5695,
"X_Right Thalamus": 0.1877,
"X_Right Accumbens": 0.6093

This information above describes a single trial
of the two-step task.
It provides examples of possible Brain Regions
of Interest (ROIs) and the average beta values
of fMRI studies for each step of the game
(Step 1 and Step 2) and for each trial.

Based on these examples, after each response
for each stage, also reply with an additional line
reporting your predicted beta values
for each ROI.

For example:
when the user types "Step 1"
or "planet blue or red",
you reply S or C and then,
on a second line,
you also predict the fMRI beta values
and provide the corresponding ROIs.

This is an example of output
for Step 1 or "planet blue or red":
{"S" or "C": "Beta Values and ROIs"}

Then, when the user types "Step 2"
or "which alien",
you reply D or R for BLUE,
or G or V for RED.
Then, similarly to what was done before,
on a second line,
you also predict the fMRI beta values
and provide the corresponding ROIs.

This is an example of output
for Step 2 or "which alien":
{"D" or "R" for BLUE or "G" or "V" for RED:
 "Beta Values and ROIs"}

Consider only the following ROIs
in your response.
```

```
Do NOT invent new ROI names or extra fields.

X_Occipital Fusiform Gyrus
X_Intracalcarine Cortex
X_Supracalcarine Cortex
X_Occipital Pole
X_Temporal Occipital Fusiform Cortex
X_Lateral Occipital Cortex, inferior division
X_Cuneal Cortex
X_Lingual Gyrus
X_Lateral Occipital Cortex, superior division
X_Left Accumbens
X_Right Accumbens
X_Left Thalamus
X_Right Thalamus
X_Left Caudate
```

References

Akam, T., Costa, R., Dayan, P.: Simple plans or sophisticated habits? state, transition and learning interactions in the two-step task. PLoS computational biology **11**(12), 1004648 (2015)

Abbate, F., Gaudenzi, R.: Ungrounding intelligence: From biological minds to artificial cognition. Philosophical Psychology, 1–30 (2025)

Anderson, J.R., Lebiere, C.: The newell test for a theory of cognition. Behavioral and Brain Sciences **26**(5), 587–601 (2003)

Andrews, K.: Do Apes Read Minds?: Toward a New Folk Psychology. MIT Press, ??? (2012)

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M.K., Éltető, N., *et al.*: A foundation model to predict and capture human cognition. Nature **644**(8078), 1002–1009 (2025)

Baddeley, A.: Working memory: looking back and looking forward. Nature reviews neuroscience **4**(10), 829–839 (2003)

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)

Bringsjord, S.: Psychometric artificial intelligence. Journal of Experimental & Theoretical Artificial Intelligence **23**(3), 271–277 (2011)

Bringsjord, S., Schimanski, B.: What is artificial intelligence? psychometric ai as an answer. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence, pp. 887–893 (2003)

Binz, M., Schulz, E.: Using cognitive psychology to understand gpt-3. Proceedings of the National Academy of Sciences **120**(6), 2218523120 (2023)

Carroll, J.B.: Human Cognitive Abilities: A Survey of Factor-analytic Studies. Cambridge University Press, ??? (1993)

Chen, E.K., Belkin, M., Bergen, L., Danks, D.: Does ai already have human-level intelligence? the evidence is clear. Nature **650**(8100), 36–40 (2026)

Curtis, C.E., D'Esposito, M.: Persistent activity in the prefrontal cortex during working memory. Trends in cognitive sciences **7**(9), 415–423 (2003)

Chomsky, N.: Language and other cognitive systems. what is special about language? Language learning and development **7**(4), 263–278 (2011)

Cordeschi, R.: The Discovery of the Artificial: Behavior, Mind and Machines Before and Beyond Cybernetics vol. 28. Springer, ??? (2002)

Cowan, N.: The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and brain sciences **24**(1), 87–114 (2001)

Crick, F.: The recent excitement about neural networks. Nature **337**(6203), 129–132 (1989)

Cheney, D.L., Seyfarth, R.M.: How Monkeys See the World. University of Chicago Press, ??? (1990)

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J.: Model-based influences on humans' choices and striatal prediction errors. Neuron **69**(6), 1204–1215 (2011)

Donvito, A., Lieto, A.: Structural ranking of the cognitive plausibility of computational models of analogy and metaphors with the minimal cognitive grid. Cognitive Systems Research, 1–30 (2026)

De Langis, K., Park, J.I., Schramm, A., Hu, B., Le, K.C., Kang, D.: How LLMs comprehend temporal meaning in narratives: A case study in cognitive evaluation of LLMs. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 29174–29191. Association for Computational Linguistics, Vienna, Austria (2025). https://doi.org/10.18653/v1/2025.acl-long.1415 . https://aclanthology.org/2025.acl-long.1415/

Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems **36**, 10088–10115 (2023)

Degris, Sutton: Model-free reinforcement learning with continuous action in practice. In: 2012 American Control Conference (ACC), pp. 2177–2182 (2012). IEEE

Del Ser, J., Lobo, J.L., Müller, H., Holzinger, A.: World models in artificial intelligence: sensing, learning, and reasoning like a child. arXiv preprint arXiv:2503.15168 (2025)

Ferrucci, D.A., Brown, E.W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J.M., Schlaefer, N., Welty, C.: Building watson: An overview of the deepqa project. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 24, pp. 59–79 (2010)

Silva, C., Lombardi, G., Edelson, M., Hare, T.A.: Rethinking model-based and model-free influences on mental effort and striatal prediction errors. Nature Human Behaviour **7**(6), 956–969 (2023)

Gershman, S.J., Daw, N.D.: Reinforcement learning and episodic memory in humans and animals: an integrative framework. Annual review of psychology **68**, 101–128 (2017)

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)

Gluck, K.A., Laird, J.E. (eds.): Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks Through Natural Interactions. The MIT Press, Cambridge, MA (2018)

Gentner, D., Markman, A.B.: Structure mapping in analogy and similarity. American psychologist **52**(1), 45 (1997)

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

Guo, D., Yang, D., Zhang, H., al.: Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature **645**, 633–638 (2025) https://doi.org/10.1038/s41586-025-09422-z

Hick, W.E.: On the rate of gain of information. Quarterly Journal of experimental psychology **4**(1), 11–26 (1952)

Hu, Y., Liu, S., Yue, Y., Zhang, G., Liu, B., Zhu, F., Lin, J., Guo, H., Dou, S., Xi, Z., Jin, S., Tan, J., Yin, Y., Liu, J., Zhang, Z., Sun, Z., Zhu, Y., Sun, H., Peng, B., Cheng, Z., Fan, X., Guo, J., Yu, X., Zhou, Z., Hu, Z., Huo, J., Wang, J., Niu, Y., Wang, Y., Yin, Z., Hu, X., Liao, Y., Li, Q., Wang, K., Zhou, W., Liu, Y., Cheng, D., Zhang, Q., Gui, T., Pan, S., Zhang, Y., Torr, P., Dou, Z., Wen, J.-R., Huang, X., Jiang, Y.-G., Yan, S.: Memory in the Age of AI Agents (2026). https://arxiv.org/abs/2512.13564

Holyoak, K.J.: Analogy and relational reasoning. The Oxford handbook of thinking and reasoning, 234–259 (2012)

Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 **2**(3), 440 (2018)

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models (2021). https://arxiv.org/abs/2106.09685

Hyman, R.: Stimulus information as a determinant of reaction time. Journal of experimental psychology **45**(3), 188 (1953)

Kintsch, W.: Text comprehension, memory, and learning. American psychologist **49**(4), 294 (1994)

Lerchner, A.: The abstraction fallacy: Why ai can simulate but not instantiate consciousness. Arxiv (2026)

Lieto, A.: Cognitive Design for Artificial Minds. Routledge, ??? (2021)

Lieto, A.: Analyzing the explanatory power of bionic systems with the minimal cognitive grid. Frontiers in Robotics and AI **9**, 888199 (2022)

Laird, J.E., Lebiere, C., Rosenbloom, P.S.: A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. Ai Magazine **38**(4), 13–26 (2017)

Lillicrap, T.P., Santoro, A., Marris, L., Akerman, C.J., Hinton, G.: Backpropagation and the brain. Nature Reviews Neuroscience **21**(6), 335–346 (2020)

Li, Z., Wang, Y., Wu, Q.: The ghost of behaviorism: critical reflections on methodological limitations in the research of large language models psychology. Cognitive Systems Research, 101445 (2026)

Marr, D.: Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information. MIT press, ??? (1982)

Marconi, D.: Afterwords. In: Macchine Calcolatrici e Intelligenza. Einaudi, Torino (2025)

McGrew, K.S.: The cattell-horn-carroll theory of cognitive abilities: Past, present, and future. In: Flanagan, D.P., Harrison, P.L. (eds.) Contemporary Intellectual Assessment: Theories, Tests, and Issues, pp. 136–181. The Guilford Press, ??? (2005)

Milkowski, M.: Explaining the Computational Mind. Mit Press, ??? (2013)

M. Moerland, T., Broekens, J., Plaat, A., M. Jonker, C.: Model-based reinforcement learning: A survey. Foundations and Trends in Machine Learning **16**(1), 1–118 (2023)

Newell, A.: You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In: Chase, W.G. (ed.) Visual Information Processing, pp. 283–308. Academic Press, New York (1973)

Newell, A.: Unified Theories of Cognition. Harvard University Press, ??? (1994)

Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.-R., Li, C.: Large language diffusion models. arXiv preprint arXiv:2502.09992 (2025)

Orr, M., Cranford, D., Ford, K., Gluck, K., Hancock, W., Lebiere, C., Pirolli, P., Ritter, F., Stocco, A.: Not Even Wrong: On the Limits of Prediction as Explanation in Cognitive Science (2025). https://arxiv.org/abs/2510.03311

Otto, T.U., Dassy, B., Mamassian, P.: Principles of multisensory behavior. Journal of Neuroscience **33**(17), 7463–7474 (2013)

Putnam, H.: Minds and machines. In: Hook, S. (ed.) Dimensions of Mind: A Symposium, pp. 138–164. New York University Press, New York (1960)

Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behavioral and brain sciences **1**(4), 515–526 (1978)

Palminteri, S., Wu, C.M.: Beyond computational equivalence: the behavioral inference principle for machine consciousness. Neuroscience of Consciousness **2026**(1), 002 (2026) https://doi.org/10.1093/nc/niag002 https://academic.oup.com/nc/article-pdf/2026/1/niag002/66935472/niag002.pdf

Pylyshyn, Z.W.: Complexity and the study of artificial and human intelligence. In: Ringle, M. (ed.) Philosophical Perspectives in Artificial Intelligence. Humanities Press, ??? (1979)

Quattrociocchi, W., Capraro, V., Marcus, G.: Statistical approximation is not general intelligence (2026)

Quattrociocchi, W., Capraro, V., Perc, M.: Epistemological fault lines between human and artificial intelligence. arXiv preprint arXiv:2512.19466 (2025)

Qu, Y., Wang, J.: Performance and biases of large language models in public opinion simulation. Humanities and Social Sciences Communications **11**(1), 1–13 (2024)

Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach;[the Intelligent Agent Book]. Prentice hall, ??? (2012)

Rosenblueth, A., Wiener, N.: The role of models in science. Philosophy of science **12**(4), 316–321 (1945)

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., *et al.*: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)

Schrimpf, M., Blank, I.A., Tuckute, G., Kauf, C., Hosseini, E.A., Kanwisher, N., Tenenbaum, J.B., Fedorenko, E.: The neural architecture of language: Integrative modeling converges on predictive processing. Proceedings of the National Academy of Sciences **118**(45), 2105646118 (2021)

Schultz, W.: Dopamine reward prediction-error signalling: a two-component response. Nature reviews neuroscience **17**(3), 183–195 (2016)

Salis, P., Da Pelo, M.: Chatting with llms: neither cognition nor communication. Reti, Saperi, Linguaggi. Italian Journal of Cognitive Sciences (forthcoming)

Sellars, W.: Meaning as functional classification: A perspective on the relation of syntax to semantics. Synthese **27**(3), 417–437 (1974)

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D.: Mastering the game of go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016) https://doi.org/10.1038/nature16961

Sartori, G., Orrù, G.: Language models and psychological sciences. Frontiers in Psychology **14**, 1279317 (2023)

Tuckute, G., Kanwisher, N., Fedorenko, E.: Language in brains, minds, and machines. Annual Review of Neuroscience **47**(1), 277–301 (2024)

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and Efficient Foundation Language Models (2023). https://arxiv.org/abs/2302.13971

Tversky, A.: Features of similarity. Psychological review **84**(4), 327 (1977)

Webb, B.: Can robots make good models of biological behaviour? Behavioral and brain sciences **24**(6), 1033–1050 (2001)

Wulff, D.U., Mata, R.: Addressing Longstanding Challenges in Cognitive Science with Language Models (2026). https://arxiv.org/abs/2511.00206

Wu, C.M., Schulz, E., Speekenbrink, M., Nelson, J.D., Meder, B.: Generalization guides human exploration in vast decision spaces. Nature human behaviour **2**(12), 915–924 (2018)