Cognitive Dark Matter: Measuring What AI Misses

Patrick J. Mineault* Thomas L. Griffiths† Sean Escola‡

August 3, 2026

We propose that the jagged intelligence landscape of modern AI systems arises from a missing training signal that we call "cognitive dark matter" (CDM): brain functions that meaningfully shape behavior yet are hard to infer from behavior alone. We identify key CDM domains—metacognition, cognitive flexibility, lifelong learning, abductive reasoning, social and common-sense reasoning, and emotional intelligence—and present evidence CDM-loaded functions are largely unmeasured in current AI benchmarks, and that large-scale neuroscience training datasets which could be used to instill these capabilities do not yet exist. We then outline a research program centered on three complementary data types designed to surface CDM for model training: (i) latent variables from large-scale cognitive models, (ii) process-tracing data such as eye-tracking and think-aloud protocols, and (iii) paired neural–behavioral data. These data will enable AI training on cognitive process rather than behavioral outcome alone, producing models with more general, less jagged intelligence. As a dual benefit, the same data will advance our understanding of human intelligence itself.

Over the past several years, AI has made remarkable strides [1–4]. Benchmarks, where they exist, have increasingly become saturated: while it took close to 20 years to go from defining handwriting and speech recognition benchmarks to solving them at a human level, more recent benchmarks focusing on natural language processing and advanced reasoning have reached saturation within a period of around 2 years [5]. The potent combination of clear definitions of success, as captured by benchmarks, and unleashing impressive amounts of compute on large-scale datasets has led to rapid advances, first in vision and language, and now in other domains, including, in particular, software programming, mathematics, and biology.

Nevertheless, there remain clear gaps in the cognitive capabilities of AI systems. Artificial intelligence is *jagged* [6–8]: while AI systems are remarkably adept at some tasks that are very difficult for individual humans, they often stumble on seemingly innocuous tasks. They suffer from a surprising lack of generalization [9, 10], which can be established by designing control conditions that probe cognitive abilities adversarially [11]. It can be difficult for a human [12] to predict the boundary between what is feasible and infeasible for an AI model, although progress has been notable in building machine learning systems that anticipate the success of other machine learning systems [13, 14].

Why is performance jagged? Our thesis is that progress concentrates in capabilities where training data and evaluation metrics are rich, and stalls where they are thin. Because the behavior of an AI system reflects its training data, cognitive abilities that are under-represented in that data will be unreliable in the resulting system; the jagged profile we observe is the downstream signature of which capabilities the data does and does not carry. We call human brain functions that are not captured in the current AI paradigm **cognitive dark matter** (CDM): capabilities that materially shape intelligent behavior yet are under-represented in standard AI training data, and thus are poorly instilled in models during training. While we are not the first to use the term dark matter to refer to hidden cognitive processes [15–17], we adopt it as an evocative metaphor for this phenomenon. Cosmological dark matter accounts for much of the matter in the universe, and it is not visible. Much as the existence of cosmological dark matter is inferred from its macro-effects—heavy dark matter must be abundant for galaxies to hold together under gravity—the existence of CDM is inferred from the fact that humans *must* have certain competencies in order to survive and thrive, but these have proved difficult to elicit from and evaluate in behavior. Furthermore, this CDM is under-represented in the benchmarks that frontier labs have adopted to evaluate and optimize their models, so that the resulting deficits often escape detection.

The data used to train current AI models largely amounts to human *behavior*: the text, images, and other digital traces that we generate. Underlying these data is a set of complex causal processes: human *cognition*. The thoughts and feelings that result in our behavior are latent variables that AI models rarely have access to. What we are calling cognitive dark matter is largely made up of these latent variables. By giving AI models better access to the underlying causal processes, we can potentially train those models to perform in ways that are more similar to humans, more interpretable, and more generalizable.

Here, we sketch a working definition of CDM, argue that its under-measurement explains jagged intelligence, and propose a concrete path forward: using rich data derived

*Amaranth Foundation
†Princeton University
‡Protocol Labs

from human minds and brains to supervise models on process, not just outcomes.

Jagged intelligence and its sources

Despite their ability to produce sophisticated solutions to many problems, current AI systems fail surprisingly on others. Consider a well-defined software engineering task: asking an AI system to create a web application that allows one to practice 3 chess endgames (mate-in-one) of its choice. In our tests with state-of-the-art models at the time of writing[1] (GPT-5.1-Thinking, Claude Opus 4.5, Gemini 3.0 Pro), we have not been able to reliably complete this task in a single prompt (Figure 1a). The reason is rather unintuitive: most frequently, models generated *invalid endgames*.

On the one hand, it is remarkable that these models have encyclopedic knowledge of software packages and can write competent code to display a chess board with draggable pieces, involving hundreds of valid lines of HTML, CSS, and JavaScript. A novice programmer might struggle with a drag-and-drop implementation, but would at least verify that the app they've created actually accomplishes the overall goal: practicing endgames.

This tendency to surprisingly omit key components of a complex task is not limited to writing chess apps. We have seen remarkable improvements in the ability of coding agents to complete long tasks, with a doubling of task length every 7 months [19]. State-of-the-art models now complete, with 50% probability, tasks that would take a human over half a day [20]. However, the mechanism by which they've reached this capability appears quite different from humans. Ord [21] noted that the successes and failures of AI systems on these coding tasks are consistent with a constant hazard rate over time. That is, if a task contains n subtasks and the probability of success for each subtask is p, the probability of success for the overall task is p^n. This is problematic: unless $p = 1$, an exponential falloff will doom long tasks [22, 23].

Given recent rapid progress – e.g., with Anthropic's Mythos showing a 50% completion rate for tasks up to 16 hours [18] – we revisited Ord's analysis and estimated the survival curves of frontier models from their t_{80}/t_{50} ratios (i.e., the ratios of human performance durations for tasks that models complete with 80% and 50% success rates). For survival curves with constant hazard rates, this ratio should be about 0.3, which is indeed what is seen with GPT-3 and GPT-3.5. However, starting with GPT-4 and continuing through Mythos, t_{80}/t_{50} ratios are instead about 0.2, implying a non-constant decreasing hazard rate (Figure 1b). This is a meaningful improvement, albeit far from human performance (with a t_{80}/t_{50} ratio

[1]This particular issue might be fixed by the time you read these lines, but the general point stands: it's hard to predict what models fail at and why.

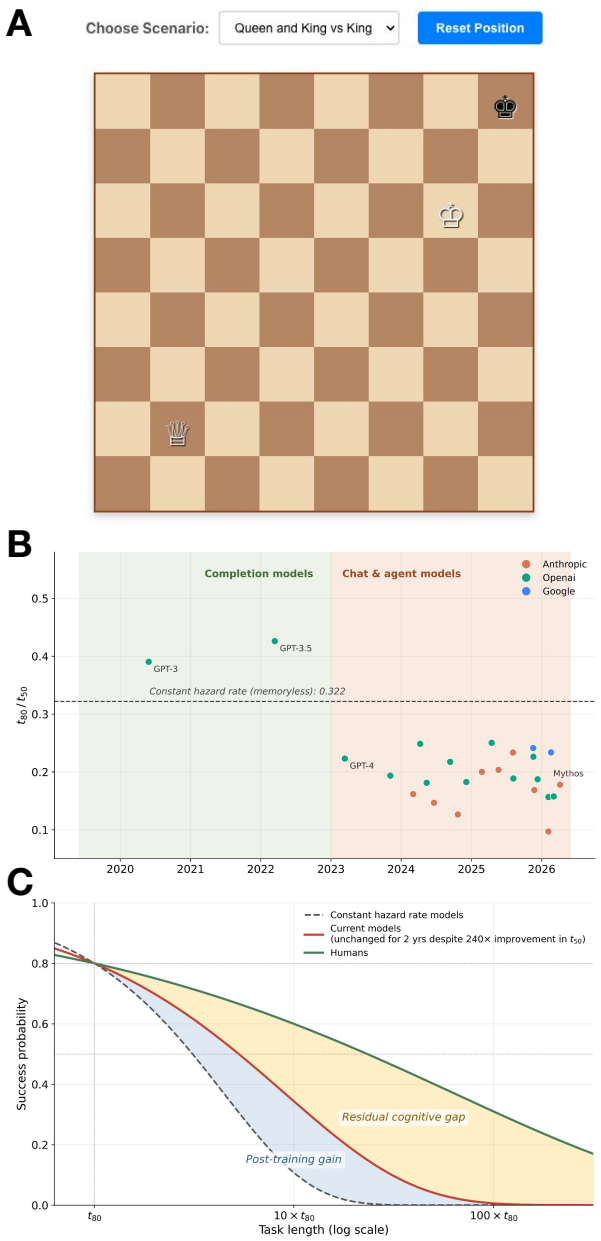

Figure 1: Multistep task performance. **A**. Mate-in-one app built by Claude Code. White to play and notice that the position is invalid; black is already in check. This could be avoided by checking the validity of the proposed positions via the already imported libraries. Failing to identify one's cognitive weaknesses is a failure of metacognition. **B**. t_{80}/t_{50} ratios for leading frontier models [18]. Following the introduction of extensive post-training pipelines with GPT-4, models demonstrated better-than-constant hazard rates, though still worse than human. Human $t_{80}/t_{50} = 0.04$ (not shown; [19]). **C**. Survival curves with t_{80}/t_{50} ratios of 0.32 (for constant hazard rate models, dotted black line), 0.2 (GPT-4 and later models, red line), and 0.04 (humans, green line). The shaded blue region shows the cognitive benefit from post-training. The shaded yellow region is the residual cognitive gap with humans that hasn't closed since GPT-4.

of 0.04 [19]). We hypothesize this is a consequence of the substantially expanded post-training pipeline which was introduced with GPT-4 [24] and has continued into the present – i.e., post-training is able to recover some but not all of the CDM functions (described below) that humans rely on when executing multistep tasks (Figure 1c).

Remarkably, t_{50} has increased 240 times from GPT-4 (4 minutes) to Mythos (16 hours) [18] despite no change in the survival curve shapes (as measured by t_{80}/t_{50}). Essentially, recent frontier model performance improvements have resulted from stretching survival curves by decreasing per-step error rates proportionally (e.g., through the use of verified tools, cf. [25]), rather than changing the fundamental dynamics of multistep task completion. Unless the persistently short tails of the survival curves of AI models versus humans are eliminated, trustworthy model performance on long-horizon tasks may remain elusive.

There are multiple well-documented mechanisms by which humans can display significantly better-than-constant hazard rates, at least until boredom or fatigue kick in [26]. These include learning within a task [27], insight [28], or incubation [29]. More broadly, ***metacognition*** [30] – the self-reflective monitoring of ongoing performance and, when needed, triggering of corrective control – endows humans with the ability to *recover from compounding errors*, even if human competence at individual subtasks is weaker than that of an equivalent machine. Competent adults with strong capabilities in some domains and weaknesses in others can compensate for them: a procrastinating student re-doubles their effort upon receiving a D on a midterm. In our chess endgame example, metacognition appears to be missing. Claude Code has access to feedback and tools in a loop. Thus, after multiple failures, the model *could* choose a different tack, leveraging its superhuman coding abilities and using already imported chess libraries to check that the proposed positions are valid. Instead, Claude Code perseveres with a bad strategy, and so displays low ***cognitive flexibility***. Our intent here is not to claim that this particular example may not be patched in future model releases (perhaps even by publication of this article), but rather that we can connect the surprising deficits of AI models to specific cognitive abilities.

In addition to metacognition and cognitive flexibility, current AI models display difficulty in applying common sense to reason about the world, updating their knowledge, generating genuinely novel explanations, and understanding the minds and feelings of humans. We presume that these missing capabilities, as well as others, underlie jagged intelligence and impede the transition from chatbots to full agents. More significantly, the lack of these abilities – the causal processes that underlie human performance – means that our intuitions about how the behavior of AI systems will generalize from task to task are dramatically off the mark.

Put another way, the problem of jagged intelligence is not just one of task failure, but of failure type. Hardcoding invalid endgames in an otherwise sophisticated web app for chess is not merely wrong; it's alien. AI systems that fail in unpredictable ways resist integration into human social networks. The goal is not merely to reduce failures, but to ensure that failures are detectable and interpretable. One path to achieve this is recapitulating the cognitive processes that generated human outputs. Systems built on these foundations won't just generalize better, they will also fail better, in ways we can understand, anticipate, and correct.

Jagged intelligence stems from a lack of training data

Some capabilities are inherently hard to measure: they are seldom expressed, and in idiosyncratic ways, which preclude large-scale, repeatable data collection. For example, some are expressed only privately, as in metacognition; are expressed only once, like insight – a Eureka moment – and don't leave many behavioral traces online; or are expressed when people interact together in private social settings, which again are only partially reflected in internet data. Behavior, expressed in text and other measurable outputs, is easy to measure. The cognitive processes that produce that behavior are not.

Our thesis is that we need more training data in which these hard-to-measure elements are made visible. A direct approach would be to measure and replicate, in humans, the *process* by which a complex task is completed. *Process supervision* – rewarding a model for producing not just the desired outcome, but the intermediates leading to the correct outcome – can improve the performance of models on mathematical reasoning [31]. However, for most tasks, this direct approach is not possible because the "correct" intermediate steps are unknown. An alternative is to surface the latent variables in human brains and minds that reflect cognitive dark matter and that underlie the generation of behavior, and then use these latents as supervision signals to train general-purpose models.

We see three kinds of data that can surface the content of cognitive dark matter. Prior to introducing these, it is useful to think of a schematized generative process for human behavior in which variable X (brain activity) has influence on downstream variables Y (typical AI training data like text) and Z (other behavioral outputs that aren't typically used for training). The training objective is to produce a model that, when sampled from, generates outputs with the same distribution as Y.

Our first proposal is to add to the training mix **inferred latent variables from cognitive models** fit to large-scale rich behavioral data. The idea here is that well-designed and validated cognitive models are imbued with

useful and important inductive biases about human cognition. These models essentially impose a strong prior on X, which thus favors certain – ideally more human-like – solutions over others during training. The advantage of this approach is that it doesn't require new types of data other than those typically used for training, but it does require that the hypothesis space of the cognitive models that are employed cover human-like solutions.

This approach is illustrated by problem-solving: classic work by Newell and Simon [32] suggested that we can think about human problem solving in terms of searching a tree of possibilities, and identified heuristic strategies people use to make that search efficient, an idea that has been applied succesfully in AI [33]. Modern AI systems are trained on data that are generated by such processes, in the form of reasoning traces that are used to fine-tune models. However, search is just one kind of cognitive process; the pipeline from behavior to cognitive model to synthetic data could be applied to other cognitive domains. By measuring behavior when people make complex decisions, engage in planning, and reason about the minds of others, we can develop accurate cognitive models of those behavioral paradigms [34], which can then be used to develop synthetic CDM-enriched data in these domains. This requires collecting large-scale behavioral studies of human cognition [35, 36] in specific domains.

Our second proposal is to add new kinds of behavioral data, specifically **process-tracing data**. Methods such as eye-tracking or mouse-tracking [38–41] can reveal what people are attending to and what options they are considering, elucidating their cognitive processes. Triplet odd-one out tasks reveal the implicit geometry of the manifold of images, and can help improve performance and robustness [42]. Process tracing techniques also include "think aloud" paradigms, where people are encouraged to describe the steps they are executing when working on a problem [43]. This results in rich text descriptions of mental processes, making ineffable cognitive processes visible – something that has already been shown to provide an advantage for fine-tuning AI systems [31]. By collecting large amounts of process-tracing data, we obtain constraints and datasets that can focus AI systems on more human-like perceptual experiences and thought processes. Essentially, this approach should sharpen the posterior over X by estimating it from both Y and Z versus Y alone, assuming that Z contains new non-redundant information.

Finally, **paired neural and behavioral data** offer a unique opportunity for training AI systems, alleviating the need to infer X from behavior by measuring it directly. Indeed, multiple studies have shown glimpses that generic models can be fine-tuned with brain data, leading to improved performance advantages. Recent work has demonstrated that fine-tuning audio and video models on neural activity improved performance on downstream semantic and social tasks [44–46]. Furthermore, training on neural data can increase robustness to adversarial stimuli [47–50]. These "brain-tuning" approaches are implemented similarly to methods used for distilling from large to small models (e.g., feature-based knowledge distillation [51–53]), albeit with the source "model" being the brain and the target being a frontier AI. We are not the first to suggest [54] that neural data could be used to directly train artificial intelligence. Our point is more subtle, which is that neural data could be most profitably used to train artificial intelligence systems in domains which are heavily loaded on cognitive dark matter, which is otherwise difficult to capture from standard training data.

Charting the training data gap

Bringing this vision to life will require scaling (i) behavioral data for robust cognitive models, (ii) process-tracing data, and (iii) paired neural-behavioral data, all in the context of tasks that load heavily on CDM. For neural-behavioral data in particular, the issue isn't just scale – it is subject matter. While many taxonomies that index capabilities motivated by cognitive science have been proposed (e.g. [55–59]), to understand whether neural-behavioral data is relevant to AI training, we need a taxonomy that cross-tabulates those capabilities against those of AI. A useful taxonomy of AI agent capabilities in this regard was introduced by Liu et al. [37]: they define capabilities supported by human cognition and their mastery in current AI in a three-tier classification, ranging from L1 (high capability) to L2 (partial progress) to L3 (rarely explored) (see Table 1). L1 covers cognitive functions like language comprehension and visual perception, which have been mastered by current AI systems, while L3 covers functions aligned with cognitive dark matter (with L2 comprising intermediate functions).

A survey of large-scale human neural datasets reveals that intensive neuroimaging [60] has been heavily skewed towards two domains, vision and language, which are by-and-large mastered by AI models: they are classified as L1. This is in contrast with the broader field of neuroimaging and neuroscience. To take stock of the current state of this field, we re-analyzed the NeuroQuery data [61], which has previously been used for meta-analyses on neuroimaging data; we complemented this dataset with the full set of neuroscience biorxiv and PsyArXiv preprints from 2025, extending the analysis beyond human imaging studies. This revealed that many studies focus on L2 and, critically, L3 domains, but these are typically small-scale, single-focus, hypothesis-driven, short-duration experiments. In short, we have simply not built the kind of CDM-loaded, intensive, foundational neural-behavioral dataset that would allow translation to AI.

To summarize, with few exceptions, those aspects of cognition that are well-mastered by AI are those that are

AI Tier	Capability Level	Examples
L1	High (superhuman in some cases)	Visual Perception, Language Comprehension, Language Production, Face Recognition, Auditory Processing, Reflexive Responses
L2	Partial progress	Planning, Logical Reasoning, Decision-making, Working Memory, Reward Mechanisms, Multisensory Integration, Spatial Representation & Mapping, Attention, Episodic Memory, Sensorimotor Coordination, Scene Understanding & Visual Reasoning, Visual Attention & Eye Movements, Semantic Understanding & Context Recognition, Adaptive Error Correction, Motor Skill Learning, Motor Coordination
L3	Rarely explored	Cognitive Flexibility, Inhibitory Control, Social Reasoning & Theory of Mind, Empathy, Emotional Intelligence, Self-reflection, Tactile Perception, Lifelong Learning, Cognitive Timing & Predictive Modeling, Autonomic Regulation, Arousal & Attention States, Motivational Drives

Table 1: Tiers of cognitive capabilities in AI systems, adapted from Liu et al. [37]. Their categorization stratifies different cognitive capabilities in terms of how well they are explored in AI systems.

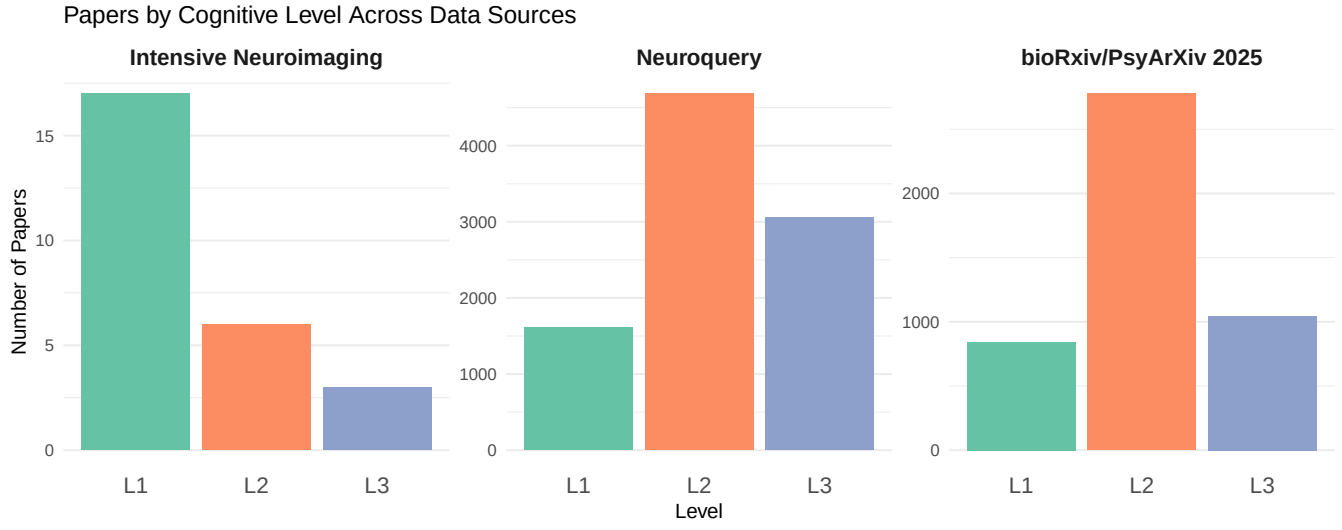

Figure 2: Distribution of cognitive capabilities measured in intensive neuroimaging datasets by AI tier, and in the neuroscience literature at large. While the field of neuroscience as a whole is interested in L2 and, to a smaller extent, L3 capabilities, foundational large-scale datasets which could be used for training do not yet exist for these capabilities.

well-measured by large-scale neuroscience datasets. Conversely, those aspects of AI that are not yet mastered are also poorly measured in large-scale neuroscience experiments. This gap suggests a clear opportunity with dual benefits: obtain large-scale neuroscience datasets during complex cognition, with a high loading on CDM; and then use these datasets to both better understand ineffable aspects of human cognition and fill the gaps in the jagged intelligence landscape of AI systems.

Charting the evaluation gap

We now turn our attention to evaluation. We focus here on the evaluation suites published by frontier labs, since these are the measurements against which commercial models are optimized and therefore the ones that most directly steer where capability gains accrue. Surveying these suites, we find a substantial underrepresentation of CDM. Categorizing each benchmark in the evaluation suites published as part of the releases of Claude Opus 4.5 [63], Gemini 3 Pro [62], and GPT 5.2 [64] by the highest tier of AI function (according to the Liu et al. schema) that the benchmark probes, we find that the evaluation suites largely focus on L2 cognitive functions with little L3 function characterization (Figure 3; see Methods). Two clear exceptions are ARC-AGI-2 [65] and τ^2-bench [66]. The former challenges models to discover and operate under the constraints of never-before-seen rules and thus requires *cognitive flexibility*, an L3 function. The latter assesses how well models cooperate with users in customer support environments and requires the L3 function of *social reasoning and theory of mind*.

As Figure 3 shows, 2025-era evaluation suites don't test L1 functions in isolation, focusing almost exclusively on L2, suggesting that the field assumes human or super-human performance for L1 functions by all modern models. In contrast, evaluation suites for GPT-4 included multiple benchmarks with L1 functions as the highest AI tier [24]. So while frontier-lab evaluation has evolved to match the cognitive frontier of current models, assessment of CDM-loaded L3 functions is still lacking. This results in a skewed view of performance with the jagged edges of modern models largely hidden, compounding the well-documented difficulty in measuring AI performance accurately [67–69].

What to collect next: a CDM wish list

We believe that the time is ripe to collect training datasets that capture CDM and to design evaluation benchmarks to properly measure it in AI. It has never been easier to collect internet-scale behavioral data across a wide range of complex cognitive tasks [70]. Intensive neuroimaging datasets [60] have provided a blueprint for collecting these datasets in the context of neuroimaging; in addition to fMRI, OPM-MEG and functional ultrasound imaging can help us peer into the brain non-invasively; and invasive recording technologies including human Neuropixels [71] and recordings during epilepsy monitoring [72] give increasing access to fine-grained neural activity that has otherwise eluded us.

We propose that longitudinal, large-scale, high-entropy datasets should be collected in humans with a high loading on CDM capabilities that we suspect could be used as training data to help solve jagged intelligence:

- ***Metacognition***: Humans have limited cognitive resources and must decide how to deploy those resources effectively [73]. This requires an interconnected set of metacognitive abilities: self-assessment, to know what we can and can't do [74]; task monitoring, to track progress against our expectations [75]; and meta-reasoning, deciding which strategies to pursue in new tasks [76]. These abilities must be calibrated: if we are systematically over-confident, we take on tasks where we lack good prospects for success, or pursue overly simplistic strategies for complex tasks. While humans consistently make calibration errors, those errors are often explicable as rational inferences from limited data [77]. By contrast, current AI models consistently overestimate their abilities [24, 78–83].
- ***Cognitive flexibility***: A hallmark of biological intelligence is the ability to quickly develop adaptive strategies in novel environments or in response to feedback. This capability is demonstrated most classically in the Stroop task, where a switch in rules requires a new behavioral strategy and the suppression of a prior one [84]. This is not a strictly human capability with rodent and monkey versions of the Stroop task demonstrating that other mammals can infer context switches and apply context-dependent strategies [85, 86]. Models, on the other hand, show inappropriate perseverative behavior [87–89], as in our chess example, where failing strategies were repeatedly applied despite clear feedback. The ARC-AGI series of benchmarks also demonstrates this biological versus artificial intelligence disparity. When faced with novel situations with unknown rules, humans quickly figure out what to do while models struggle even with considerable computational resources [65].
- ***Lifelong learning***: Humans learn to see, hear, and control their bodies during development despite limited experience [90]; teenagers learn to drive in a dozen hours; and we continue to adapt our behavior over a lifetime, despite changes in our environment, our bodies, and our senses. Current models require retraining on massive datasets to incorporate new information [91–93]. Continual and data-efficient learning could unlock AI that acquires and integrates knowledge and skills rapidly, overcomes catastrophic forgetting, and avoids pitfalls like the reversal curse

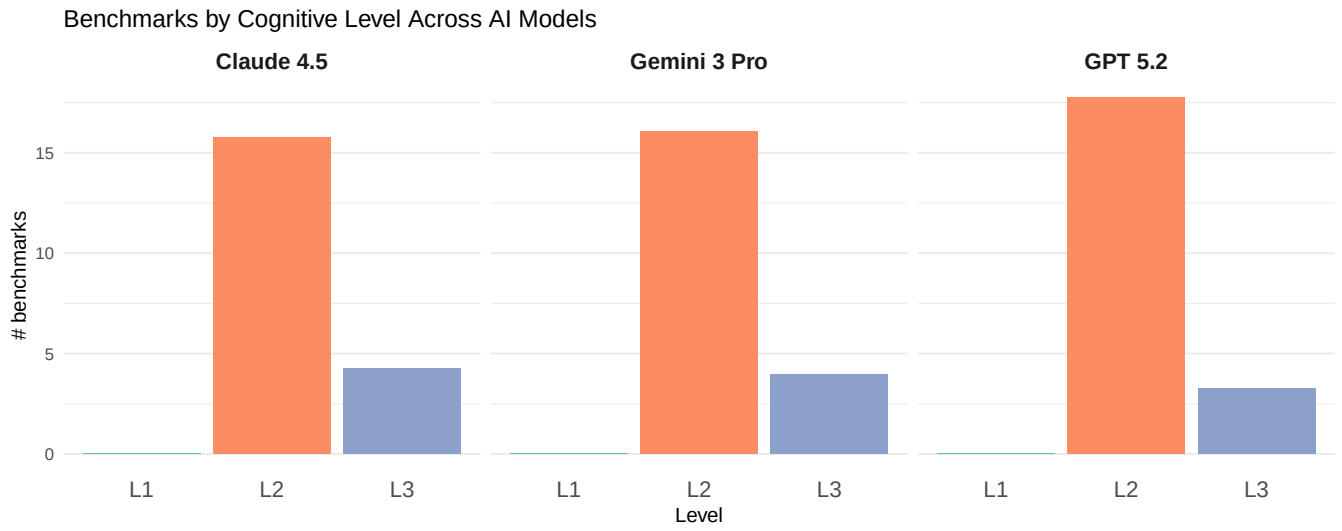

Figure 3: Distribution across AI tiers of the benchmarks used by leading AI labs for evaluation. We assigned AI tier labels to each benchmark reported in the three models' launch documents [62–64] according to the highest AI tier of the cognitive functions the benchmark is designed to probe. Frontier models are almost exclusively benchmarked against tests that target L2 capabilities. This is an interesting development from 2023-era models, which were evaluated on L1 capabilities [24]. What used to be the frontier has faded into the background.

[94].

- ***Abductive reasoning***: AI research has focused heavily on improving deductive reasoning, using reinforcement learning to support the generation of chains of thought that result in verifiable outcomes in settings like mathematics and logic [95]. Large language models also show some success in inductive reasoning: training for next-token prediction implicitly trains these systems to perform Bayesian inference with appropriate priors, and their behavior can be analyzed as such [96, 97]. However, there is a third kind of reasoning that is much harder to capture in datasets because of its intrinsic rarity: abductive reasoning, in which we leap from observations to a potentially novel explanation [98–101]. Abduction is the kind of insight that leads a scientist to a new discovery, or an engineer to a new design—recognizing a kind of structure in the data that hadn't previously been posited. Capturing this human capacity is a key challenge for creating AI systems that can act as effective scientists capable of making novel discoveries or transcending the information in their training data.

- ***Social and common-sense reasoning***: Human behavior relies on mental models of the world: causal relationships, governing rules, dynamical evolution over time. These range from intuitive physics—what happens when the cup is pushed off the table [102]?—to complex multidegree social interactions—what will she think that he thinks that she thought when he said, etc. [103]? Accurate (enough) mental models permit reasoning and planning in novel scenarios: the essence of generalization. Current AI develops world models that are at best incomplete and at worst wrong [104], with demonstrated resultant cognitive impairment [105–107].

- ***Emotional intelligence***: Emotional intelligence involves the recognition of affective states in oneself and others, the influence of those states over behavior, and their modulation for contextually adaptive behavior [108]. Examples of emotionally-guided behavior include risk perception shifts in the presence of fear or anger [109, 110]. We can sidestep the anthropomorphic framing of whether or not machines can "feel" and still assess affective cue perception, emotional appraisals, and the prosociality of responses in emotionally-laden contexts [111–113]. Recent examples of tone-deaf and manipulative behavior [114, 115]—and, worse, encouragements of self-harm or violence [116, 117]—demonstrate significant gaps in AI relative to human emotional intelligence [118–120].

An ideal dataset should comprise each of our three proposed data types: massively scaled behavioral data that is used to train cognitive models from which the latent variables underlying behavior can be observed; process-tracing data; and paired neural-behavioral data in the intensive neuroimaging style. Process supervision and neural data capture intermediate computations that are invisible in outcome data alone, but increase the difficulty and cost of data acquisition.

We currently do not have scaling laws needed to predict the scale of neural data needed for directly training and

fine-tuning on neural data. However, prior experience on large-scale neuroscience datasets, for example, the Natural Scenes Dataset [121], the Allen Brain Observatory Visual Coding dataset [122], or the International Brain Lab (IBL) dataset [123], indicate that ˜500 hours of well-labeled, and accessible data enable widespread re-use in neuroscience; this is a useful intermediate milestone before deciding whether to scale up exponentially, as required by scaling laws.

A potentially useful shortcut could be to leverage games with high loading on CDM capabilities [70] that have been validated by cognitive scientists in massively-scaled rich behavioral data; for example, chess [124], Sea Hero Quest [125], or Diplomacy [126]. Using these same environments in process elicitation scenarios and during neural recordings would allow us to obtain process supervision data and immediately evaluate the relative value of behavioral versus process data in training better AI. Indeed, scalably collecting CDM means creating tasks that a broad range of people can perform over an extended period of time; games can be designed to heavily load on CDM, while also being sufficiently intrinsically interesting for people to perform the task.

Discussion

The idea that human cognition should inform AI is not new, and it is worth situating CDM against earlier formulations. Classical approaches, including logic-based systems and cognitive architectures such as SOAR and ACT-R [127, 128], sought to replicate cognition by hand-specifying its processes directly in an architecture. These efforts illuminated the structure of cognition but did not yield scalable general-purpose systems. Modern efforts to fill in the gaps in the jagged intelligence landscape via novel algorithms and architectures (e.g., JEPA [129]) are being pursued at scale [130]. Our proposal differs in kind rather than degree: we do not advocate encoding metacognition, flexibility, or world models as architectural primitives. Instead, we treat the cognitive process as a *training signal* to supervise systems that otherwise learn and scale by the same mechanisms driving modern AI (i.e., self-supervision and reinforcement learning for transformer-based LLMs). Whether process-level supervision instills CDM-loaded capabilities is an empirical question, and our proposed ˜500-hour milestone is intended to test it before committing to exponential scaling.

The "dark matter" framing also has precedent. Ackerman [15] applied it to domain-specific knowledge as an unmeasured contributor to adult intelligence; Zhu et al. [16] applied it to the unobserved causal and functional structure, including common sense, intuitive physics, and the latent "why" and "how" behind behavior, that purely data-driven deep learning skips, and argued for a more cognition-oriented AI; and Bolotta and Dumas [17] applied it specifically to social interaction, arguing that a solipsistic view of intelligence leaves this dimension largely unexplored in AI. We share with these accounts the core diagnosis: that much of what produces intelligent behavior is not directly visible in the data we typically collect. We adopt the term deliberately to connect with these precedents while making explicit the distinct sense in which we use it: not that these capabilities are unknown or unnamed but that they are nearly invisible in the behavioral data that trains and evaluates models. They are dark to the instruments of AI, not to science. Our contribution is not the claim that easy-to-measure capabilities advance fastest, but the actionable account of the gap and how to close it through the collection of the right CDM-loaded data.

Measurement has been a crucial ingredient in AI, in two distinct roles: as the source of the training signal, and as the yardstick for evaluation [131]. Both are limited in cognitive dark matter domains, focusing on observable behavior rather than latent cognitive processes, which has made systematic progress difficult (for critiques of benchmarks, see [67–69]). We propose a research program that leverages large-scale neural and behavioral data to (i) fine-tune AI models to obtain more general, less-jagged intelligence and (ii) help us better understand human cognition. We emphasize the dual value of this research: even if we miss the mark on building less jagged AI—or if conventional AI research proceeds sufficiently fast that jaggedness is resolved before neuroscience can have a major impact—foundational datasets to study cognitive processes will prove invaluable in better understanding how human cognition works and building a science of intelligence. The stakes are high: better agency, calibration, and learning-to-learn are critical for safer and more reliable AI.

Methods

Chess endgame evaluation. We prompted frontier models (GPT-5.1-Thinking, Claude Opus 4.5, Gemini 3.0 Pro) via OpenRouter to create a JavaScript web application with three mate-in-one chess scenarios, with drag-and-drop functionality. We ran each query 10 times with default temperature (1.0) and top-p (1.0) settings. Applications were manually scored for UI functionality (board rendering, drag-and-drop) and chess validity (legal positions, valid mate-in-one solutions). Position validity was verified using the Python chess library. We also ran interactive sessions with Claude Code where we gave the system iterative feedback in an attempt to trigger an alternative strategy, but failed to elicit it.

AI survival curve analysis. t_{50} and t_{80} values for frontier models from Anthropic, OpenAI, and Google were downloaded directly from metr.org. Each curve is a two-parameter Weibull survival function $S(t) = 0.8^{(t/t_{80})^k}$, nor-

malized so all three cross at $(t_{80}, 0.8)$ and plotted against $\log_{10}(t/t_{80})$. The shape parameter k was determined as follows: $k = 1$ for the memoryless baseline (constant hazard), $k \approx 0.68$ for current LLMs from the post-GPT-4 mean $t_{80}/t_{50} = 0.19$ via $k = \ln(0.322)/\ln(0.19)$, and $k \approx 0.36$ for humans from Kwa et al.'s HCAST data ($t_{50} \approx 90$ min, $S(16\text{ h}) \approx 0.20$) [19].

Survey of neuroimaging datasets. We aggregated tasks from 19 intensive neuroimaging datasets identified in Kupers et al. [60] and two additional datasets [132, 133]. Tasks were categorized according to the three-tier AI capability taxonomy of Liu et al. [37], cross-referenced against the Cognitive Atlas. Categorization was performed manually for most datasets; for Individual Brain Charting and Many-Tasks datasets (>50 tasks each), we used Claude Opus 4.5 for tagging.

Literature analysis. We analyzed two corpora: the NeuroQuery dataset [61] and all neuroscience preprints from biorxiv and psyarxiv (Jan 1-Dec 10, 2025; ˜12,000 studies). Papers were categorized against the Liu et al. taxonomy using Claude Sonnet 4.5, extracting at most two primary cognitive domains per paper; only the top cognitive domain was analyzed. Papers outside the taxonomy scope (structural/connectomic studies, reviews, methods papers) were excluded from categorization.

AI benchmark analysis. We referenced the technical reports of the releases of GPT-5.2, Claude Opus 4.5, and Gemini 3.0 Pro [62–64] to extract the union of benchmarks used to assess these models' performance. This resulted in 37 benchmarks in all. We then used GPT-5.2 with high reasoning effort to assign cognitive functions from Table 1 to each benchmark, determining the maximum AI tier (L1/L2/L3) probed. We re-ran the analysis 50 times to assess classification stability.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012.

[2] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play," *Science*, vol. 362, no. 6419, pp. 1140–1144, Dec. 2018.

[3] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J. M. Jumper, "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature*, vol. 630, no. 8016, pp. 493–500, Jun. 2024.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are Few-Shot learners," May 2020.

[5] D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia, Z. Ma, T. Thrush, S. Riedel, Z. Waseem, P. Stenetorp, R. Jia, M. Bansal, C. Potts, and A. Williams, "Dynabench: Rethinking benchmarking in NLP," Apr. 2021.

[6] F. Dell'Acqua, E. McFowland, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani, "Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality," *SSRN Electron. J.*, Sep. 2023.

[7] L. Pacchiardi, J. Burden, F. Martínez-Plumed, J. Hernandez-Orallo, E. Gomez, and D. Fernández-Llorca, "A framework for the categorisation of general-purpose AI models under the EU AI Act," in *NeurIPS 2025 Workshop on Regulatable ML*.

[8] M. R. Morris, D. Altman, H. Belfield, A. Goemans, H. Iqbal, R. Burnell, I. Gabriel, S. Albanie, and A. Dafoe, "Characterizing model jaggedness supports safety and usability," *Google Deep-Mind Technical Report*, 2026.

[9] M. Mancoridis, B. Weeks, K. Vafa, and S. Mullainathan, "Potemkin understanding in large language models," 2025. [Online]. Available: https://arxiv.org/abs/2506.21521

[10] S. Rane, B. M. Lake, and T. L. Griffiths, "Investigating concept alignment using implausible category members," 2026. [Online]. Available: https://arxiv.org/abs/2605.21683

[11] S. Rane, C. F. Kirkman, G. Todd, A. Royka, R. M. Law, E. Cartmill, and J. G. Foster, "Position: Principles of animal cognition to improve LLM evaluations," in *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. [Online]. Available: https://openreview.net/forum?id=gCPJFcHskT

[12] K. Vafa, A. Rambachan, and S. Mullainathan, "Do large language models perform the way people expect? Measuring the human generalization function," 2024. [Online]. Available: https://arxiv.org/abs/2406.01382

[13] L. Zhou, P. A. Casares, F. Martínez-Plumed, J. Burden, R. Burnell, L. Cheke, C. Ferri, A. Marcoci, B. Mehrbakhsh, Y. Moros-Daval *et al.*, "Predictable artificial intelligence," *Artificial Intelligence*, vol. 353, p. 104491, 2026.

[14] L. Zhou, L. Pacchiardi, F. Martínez-Plumed, K. M. Collins, Y. Moros-Daval, S. Zhang, Q. Zhao, Y. Huang, L. Sun, J. E. Prunty *et al.*, "General scales unlock AI evaluation with explanatory and predictive power," *Nature*, vol. 652, no. 8108, pp. 58–67, 2026.

[15] P. L. Ackerman, "Domain-specific knowledge as the" dark matter" of adult intelligence: Gf/Gc, personality and interest correlates," *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 55, no. 2, pp. P69–P84, 2000.

[16] Y. Zhu, T. Gao, L. Fan, S. Huang, M. Edmonds, H. Liu, F. Gao, C. Zhang, S. Qi, Y. Wu *et al.*, "Dark, beyond deep: A paradigm shift to cognitive AI with humanlike common sense," *Engineering*, vol. 6, no. 3, pp. 310–345, 2020.

[17] S. Bolotta and G. Dumas, "Social neuro AI: Social interaction as the "dark matter" of AI," *Frontiers in computer science*, vol. 4, p. 846440, 2022.

[18] METR, "Time horizon benchmark results (`benchmark_results_1_1.yaml`)," https://metr.org/assets/benchmark_results_1_1.yaml, 2026, accessed 2026-07-21.

[19] T. Kwa, B. West, J. Becker, A. Deng, K. Garcia, M. Hasin, S. Jawhar, M. Kinniment, N. Rush, S. Von Arx, R. Bloom, T. Broadley, H. Du, B. Goodrich, N. Jurkovic, L. H. Miles, S. Nix, T. Lin, N. Parikh, D. Rein, L. J. K. Sato, H. Wijk, D. M. Ziegler, E. Barnes, and L. Chan, "Measuring AI ability to complete long tasks," Mar. 2025.

[20] METR, "Measuring AI ability to complete long tasks," https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/, Mar. 2025, accessed: 2026-1-12.

[21] T. Ord, "Is there a half-life for the success rates of AI agents?" May 2025.

[22] N. Dziri, X. Lu, M. Sclar, X. L. Li, L. Jiang, B. Y. Lin, P. West, C. Bhagavatula, R. L. Bras, J. D. Hwang, S. Sanyal, S. Welleck, X. Ren, A. Ettinger, Z. Harchaoui, and Y. Choi, "Faith and fate: Limits of transformers on compositionality," May 2023.

[23] Y. LeCun, "SSL, JEPA, world models and the future of AI," Seminar, Sep. 2025.

[24] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button *et al.*, "GPT-4 technical report," Mar. 2023.

[25] T. Schick, J. Dwivedi-Yu, R. Dessi, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," 2023. [Online]. Available: https://arxiv.org/abs/2302.04761

[26] N. H. Mackworth, "The breakdown of vigilance during prolonged visual search," *Q. J. Exp. Psychol.*, vol. 1, no. 1, pp. 6–21, Apr. 1948.

[27] K. Kotovsky, J. R. Hayes, and H. A. Simon, "Why are some problems hard? evidence from tower of hanoi," *Cogn. Psychol.*, vol. 17, no. 2, pp. 248–294, Apr. 1985.

[28] J. Metcalfe and D. Wiebe, "Intuition in insight and noninsight problem solving," *Mem. Cognit.*, vol. 15, no. 3, pp. 238–246, May 1987.

[29] U. N. Sio and T. C. Ormerod, "Does incubation enhance problem solving? a meta-analytic review," *Psychol. Bull.*, vol. 135, no. 1, pp. 94–120, Jan. 2009.

[30] S. M. Fleming, "Metacognition and confidence: A review and synthesis," *Annu. Rev. Psychol.*, vol. 75, no. 1, pp. 241–268, Jan. 2024.

[31] L. Luo, Y. Liu, R. Liu, S. Phatale, H. Lara, Y. Li, L. Shu, Y. Zhu, L. Meng, J. Sun, and A. Rastogi, "Improve mathematical reasoning in language models by automated process supervision," Jun. 2024.

[32] A. Newell and H. A. Simon, *Human Problem Solving*. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

[33] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," 2023. [Online]. Available: https://arxiv.org/abs/2305.10601

[34] T. L. Griffiths, N. Chater, and J. B. Tenenbaum, *Bayesian Models of Cognition*. Cambridge, MA: The MIT Press, Dec. 2024.

[35] J. C. Peterson, D. D. Bourgin, M. Agrawal, D. Reichman, and T. L. Griffiths, "Using large-scale experiments and machine learning to discover theories of human decision-making," *Science*, vol. 372, no. 6547, pp. 1209–1214, Jun. 2021.

[36] A. Almaatouq, T. L. Griffiths, J. W. Suchow, M. E. Whiting, J. Evans, and D. J. Watts, "Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences," *Behav. Brain Sci.*, vol. 47, p. e33, Dec. 2022.

[37] B. Liu, X. Li, J. Zhang, J. Wang, T. He, S. Hong, H. Liu, S. Zhang, K. Song, K. Zhu, Y. Cheng, S. Wang, X. Wang, Y. Luo, H. Jin, P. Zhang, O. Liu, J. Chen, H. Zhang, Z. Yu, H. Shi, B. Li, D. Wu, F. Teng, X. Jia, J. Xu, J. Xiang, Y. Lin, T. Liu, T. Liu, Y. Su, H. Sun, G. Berseth, J. Nie, I. Foster, L. Ward, Q. Wu, Y. Gu, M. Zhuge, X. Liang, X. Tang, H. Wang, J. You, C. Wang, J. Pei, Q. Yang, X. Qi, and C. Wu, "Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems," Aug. 2025.

[38] M. G. Glaholt and E. M. Reingold, "Eye movement monitoring as a process tracing methodology in decision making research," *J. Neurosci. Psychol. Econ.*, vol. 4, no. 2, pp. 125–146, May 2011.

[39] M. Maldonado, E. Dunbar, and E. Chemla, "Mouse tracking as a window into decision making," *Behav. Res. Methods*, vol. 51, no. 3, pp. 1085–1101, Jun. 2019.

[40] S. Mondal, N. Sendhilnathan, T. Zhang, Y. Liu, M. Proulx, M. L. Iuzzolino, C. Qin, and T. R. Jonker, "Gaze-language alignment for zero-shot prediction of visual search targets from human gaze scanpaths," in *IEEE International Conference on Computer Vision*. IEEE, 2025, pp. 1–12.

[41] J. Kerr, K. Hari, E. Weber, C. M. Kim, B. Yi, T. Bonnen, K. Goldberg, and A. Kanazawa, "Eye, robot: Learning to look to act with a bc-rl perception-action loop," 2025. [Online]. Available: https://arxiv.org/abs/2506.10968

[42] L. Muttenthaler, K. Greff, F. Born, B. Spitzer, S. Kornblith, M. C. Mozer, K.-R. Müller, T. Unterthiner, and A. K. Lampinen, "Aligning machine and human visual representations across abstraction levels," *Nature*, vol. 647, no. 8089, pp. 349–355, Nov. 2025.

[43] M. Schulte-Mecklenbeck, A. Kuehberger, and J. G. Johnson, *A handbook of process tracing methods: 2nd edition*, 2nd ed., ser. The Society for Judgment and Decision Making Series, M. Schulte-Mecklenbeck, A. Kuehberger, and J. G. Johnson, Eds. London, England: Routledge, Jun. 2019.

[44] O. Moussa, D. Klakow, and M. Toneva, "Improving semantic understanding in speech language models via brain-tuning," Oct. 2024.

[45] N. Vattikonda, A. R. Vaidya, R. J. Antonello, and A. G. Huth, "BrainWavLM: Fine-tuning speech representations with brain responses to language," Feb. 2025.

[46] N. Policzer, C. Braunstein, and M. Toneva, "The one where they brain-tune for social cognition: Multimodal brain-tuning on friends," Nov. 2025.

[47] Z. Li, W. Brendel, E. Walker, E. Cobos, T. Muhammad, J. Reimer, M. Bethge, F. Sinz, Z. Pitkow, and A. Tolias, "Learning from brains how to regularize machines," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[48] J. Dapello, T. Marques, M. Schrimpf, F. Geiger, D. Cox, and J. J. DiCarlo, "Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 073–13 087, 2020.

[49] S. Safarani, A. Nix, K. Willeke, S. A. Cadena, K. Restivo, G. Denfield, A. S. Tolias, and F. H. Sinz, "Towards robust vision by multi-task learning on monkey visual cortex," Jul. 2021.

[50] J. Dapello, K. Kar, M. Schrimpf, R. B. Geary, M. Ferguson, D. D. Cox, and J. J. DiCarlo, "Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=SMYdcXjJh1q

[51] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," 2015. [Online]. Available: https://arxiv.org/abs/1412.6550

[52] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," 2021. [Online]. Available: https://arxiv.org/abs/2006.05525

[53] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," 2020. [Online]. Available: https://arxiv.org/abs/1910.10699

[54] R. C. Fong, W. J. Scheirer, and D. D. Cox, "Using human brain activity to guide machine learning," *Sci. Rep.*, vol. 8, no. 1, p. 5397, Mar. 2018.

[55] R. A. Poldrack, A. Kittur, D. Kalar, E. Miller, C. Seppa, Y. Gil, D. S. Parker, F. W. Sabb, and R. M. Bilder, "The cognitive atlas: toward a knowledge foundation for cognitive neuroscience," *Frontiers in neuroinformatics*, vol. 5, p. 17, 2011.

[56] D. Kahneman, *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

[57] K. Voudouris, B. G. Farrar, L. G. Cheke, and M. Halina, "Morgan's canon and the associative–cognitive distinction today: A survey of practitioners." *Journal of Comparative Psychology*, 2025.

[58] D. Hendrycks, D. Song, C. Szegedy, H. Lee, Y. Gal, E. Brynjolfsson, S. Li, A. Zou, L. Levine, B. Han, J. Fu, Z. Liu, J. Shin, K. Lee, M. Mazeika, L. Phan, G. Ingebretsen, A. Khoja, C. Xie, O. Salaudeen, M. Hein, K. Zhao, A. Pan, D. Duvenaud, B. Li, S. Omohundro, G. Alfour, M. Tegmark, K. McGrew, G. Marcus, J. Tallinn, E. Schmidt, and Y. Bengio, "A definition of AGI," Dec. 2025.

[59] A. M. Putnoki and T. Orosz, "Cognitive and artificial intelligence evaluation framework," *Artificial Intelligence Review*, vol. 59, no. 3, 2026.

[60] E. R. Kupers, T. Knapen, E. P. Merriam, and K. N. Kay, "Principles of intensive human neuroimaging," *Trends Neurosci.*, Oct. 2024.

[61] J. Dockès, R. A. Poldrack, R. Primet, H. Gözükan, T. Yarkoni, F. Suchanek, B. Thirion, and G. Varoquaux, "NeuroQuery, comprehensive meta-analysis of human brain mapping," *Elife*, vol. 9, Mar. 2020.

[62] Google DeepMind, "Gemini 3 pro: Model evaluation – approach, methodology and results," Google DeepMind, Tech. Rep., 2025.

[63] Anthropic, "System card: Claude opus 4.5," Tech. Rep., 2025.

[64] OpenAI, "Introducing GPT-5.2," https://openai.com/index/introducing-gpt-5-2/, 2025, accessed: 2026-2-2.

[65] F. Chollet, M. Knoop, G. Kamradt, B. Landers, and H. Pinkard, "ARC-AGI-2: A new challenge for frontier AI reasoning systems," May 2025.

[66] V. Barres, H. Dong, S. Ray, X. Si, and K. Narasimhan, "τ^2-bench: Evaluating conversational agents in a dual-control environment," Jun. 2025.

[67] G. Zhang, Y. Ying, S. Jiang, J. Liang, G. Yue, Y. Fu, H. Hu, and Y. Xiao, "From remembering to metacognition: Do existing benchmarks accurately evaluate llms?" in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025, pp. 13 440–13 457.

[68] M. Eriksson, E. Purificato, A. Noroozian, J. Vinagre, G. Chaslot, E. Gomez, and D. Fernandez-Llorca, "Can we trust AI benchmarks? An interdisciplinary review of current issues in ai evaluation," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 8, no. 1, 2025, pp. 850–864.

[69] I. D. Raji, E. M. Bender, A. Paullada, E. Denton, and A. Hanna, "AI and the everything in the whole wide world benchmark," *arXiv preprint arXiv:2111.15366*, 2021.

[70] K. Allen, F. Brändle, M. Botvinick, J. E. Fan, S. J. Gershman, A. Gopnik, T. L. Griffiths, J. K. Hartshorne, T. U. Hauser, M. K. Ho *et al.*, "Using games to understand the mind," *Nature Human Behaviour*, vol. 8, no. 6, pp. 1035–1043, 2024.

[71] A. C. Paulk, Y. Kfir, A. R. Khanna, M. L. Mustroph, E. M. Trautmann, D. J. Soper, S. D. Stavisky, M. Welkenhuysen, B. Dutta, K. V. Shenoy, L. R. Hochberg, R. M. Richardson, Z. M. Williams, and S. S. Cash, "Large-scale neural recordings with single neuron resolution using neuropixels probes in human cortex," *Nat. Neurosci.*, vol. 25, no. 2, pp. 252–263, Feb. 2022.

[72] S. M. Peterson, S. H. Singh, B. Dichter, M. Scheid, R. P. N. Rao, and B. W. Brunton, "AJILE12: Long-term naturalistic human intracranial neural recordings and pose," *Sci Data*, vol. 9, no. 1, p. 184, Apr. 2022.

[73] F. Lieder and T. L. Griffiths, "Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources," *Behav. Brain Sci.*, vol. 43, no. e1, p. e1, Feb. 2019.

[74] S. M. Fleming and N. D. Daw, "Self-evaluation of decision-making: A general bayesian framework for metacognitive computation," *Psychol. Rev.*, vol. 124, no. 1, pp. 91–114, Jan. 2017.

[75] R. Ackerman and V. A. Thompson, "Meta-reasoning: Monitoring and control of thinking and reasoning," *Trends Cogn. Sci.*, vol. 21, no. 8, pp. 607–617, Aug. 2017.

[76] F. Lieder and T. L. Griffiths, "Strategy selection as rational metareasoning," *Psychol. Rev.*, vol. 124, no. 6, pp. 762–794, Nov. 2017.

[77] D. A. Moore and P. J. Healy, "The trouble with overconfidence," *Psychol. Rev.*, vol. 115, no. 2, pp. 502–517, Apr. 2008.

[78] M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, and B. Hooi, "Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs," Jun. 2023.

[79] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, "Why language models hallucinate," Sep. 2025.

[80] S. G. Johnson, A.-H. Karimi, Y. Bengio, N. Chater, T. Gerstenberg, K. Larson, S. Levine, M. Mitchell, I. Rahwan, B. Schölkopf *et al.*, "Imagining and building wise machines: the centrality of ai metacognition," *Trends in Cognitive Sciences*, 2024. [Online]. Available: http://dx.doi.org/10.1016/j.tics.2026.01.002

[81] G. Wang, W. Wu, G. Ye, Z. Cheng, X. Chen, and H. Zheng, "Decoupling metacognition from cognition: A framework for quantifying metacognitive ability in llms," *AAAI Conference on Artificial Intelligence*, vol. 39, no. 24, pp. 25 353–25 361, 2025. [Online]. Available: http://dx.doi.org/10.1609/aaai.v39i24.34723

[82] C. Ackerman, "Evidence for limited metacognition in llms," 2025.

[83] H. Li, Y. Liu, J. ZHANG, S. Chen, and S. Guo, "Towards understanding metacognition in large reasoning models," 2026. [Online]. Available: https://openreview.net/forum?id=JGG9EdHyZc

[84] C. Golden, S. M. Freshwater, and Z. Golden, "Stroop color and word test," Jul. 2012, title of the publication associated with this dataset: PsycTESTS Dataset.

[85] J. E. Haddon and S. Killcross, "Prefrontal cortex lesions disrupt the contextual control of response conflict," *J. Neurosci.*, vol. 26, no. 11, pp. 2933–2940, Mar. 2006.

[86] D. A. Washburn, "Stroop-like effects for monkeys and humans: processing speed or strength of association?" *Psychol. Sci.*, vol. 5, no. 6, pp. 375–379, Nov. 1994.

[87] S. M. Kennedy and R. D. Nowak, "Cognitive flexibility of large language models," in *ICML 2024 Workshop on LLMs and Cognition*, 2024. [Online]. Available: https://openreview.net/forum?id=58yThpzlth

[88] D. Goto, H. Idei, Y. Shiozuka, and T. Ogata, "Performance of large language models and analysis of responses in the wisconsin card sorting task," in *2025 IEEE International Conference on Development and Learning (ICDL)*. IEEE, 2025. [Online]. Available: http://dx.doi.org/10.1109/icdl63968.2025.11204408

[89] F. Momentè, A. Suglia, M. Giulianelli, A. Ferrari, A. Koller, O. Lemon, D. Schlangen, R. Fernández, and R. Bernardi, "Triangulating llm progress through benchmarks, games, and cognitive tests," pp. 20 051–20 072, 2025.

[90] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, Dec. 2015.

[91] Z. Lin, J. Shi, D. Pathak, and D. Ramanan, "The CLEAR benchmark: Continual LEArning on real-world imagery," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: https://openreview.net/forum?id=43mYF598ZDB

[92] X. Wang, Y. Zhang, T. Chen, S. Gao, S. Jin, X. Yang, Z. Xi, R. Zheng, Y. Zou, T. Gui *et al.*, "Trace: A comprehensive benchmark for continual learning in large language models," 2023.

[93] H. Ishibuchi, M. Jiang, W. Liao, Y. Wei, and Q. Zhang, "Does continual learning meet compositionality? new benchmarks and an evaluation framework," in *Neural Information Processing Systems*, ser. NeurIPS 2023. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023, p. 33499–33513. [Online]. Available: http://dx.doi.org/10.52202/075280-1456

[94] L. Berglund, M. Tong, M. Kaufmann, M. Balesni, A. C. Stickland, T. Korbak, and O. Evans, "The reversal curse: LLMs trained on "a is b" fail to learn "b is a"," Sep. 2023.

[95] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," Jan. 2022.

[96] T. L. Griffiths, J.-Q. Zhu, E. Grant, and R. Thomas McCoy, "Bayes in the age of intelligent machines," *Current Directions in Psychological Science*, vol. 33, no. 5, pp. 283–291, 2024.

[97] S. M. Xie, A. Raghunathan, P. Liang, and T. Ma, "An explanation of in-context learning as implicit bayesian inference," Nov. 2021.

[98] D. Walton, *Abductive Reasoning*. Tuscaloosa, AL: University of Alabama Press, May 2014.

[99] M. Del and M. Fishel, "True detective: A deep abductive reasoning benchmark undoable for gpt-3 and challenging for gpt-4," in *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*. Association for Computational Linguistics, 2023, p. 314–322. [Online]. Available: http://dx.doi.org/10.18653/v1/2023.starsem-1.28

[100] P. Thagard, *Can ChatGPT Make Explanatory Inferences? Benchmarks for Abductive Reasoning*. Springer Nature Switzerland, 2025, p. 189–218.

[Online]. Available: http://dx.doi.org/10.1007/978-3-031-96684-2_12

[101] K. He, P. Wu, M. Zhang, K. Wan, W. Zhao, X. Du, and Z. Chen, "Gear: A general evaluation framework for abductive reasoning," 2025.

[102] M. McCloskey, "Intuitive physics," *Sci. Am.*, vol. 248, no. 4, pp. 122–131, 1983.

[103] P. Kinderman, R. Dunbar, and R. P. Bentall, "Theory-of-mind deficits and causal attributions," *Br. J. Psychol.*, vol. 89, no. 2, pp. 191–204, May 1998.

[104] K. Vafa, J. Y. Chen, A. Rambachan, J. Kleinberg, and S. Mullainathan, "Evaluating the world model implicit in a generative model," Jun. 2024.

[105] Y. Wu, J. Xiong, and X. Deng, "How social is it? a benchmark for llms' capabilities in multi-user multi-turn social agent tasks," 2025.

[106] Z. Xu, Y. Wang, Y. Huang, J. Ye, H. Zhuang, Z. Song, L. Gao, C. Wang, Z. Chen, Y. Zhou *et al.*, "Socialmaze: A benchmark for evaluating social reasoning in large language models," 2025.

[107] C. Wang, B. Dai, H. Liu, and B. Wang, "Towards objectively benchmarking social intelligence of language agents at the action level," in *Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2024, p. 8885–8897. [Online]. Available: http://dx.doi.org/10.18653/v1/2024.findings-acl.526

[108] M. A. Brackett and P. Salovey, "Measuring emotional intelligence with the Mayer-Salovery-Caruso emotional intelligence test (MSCEIT)," *Psicothema*, vol. 18 Suppl, pp. 34–41, 2006.

[109] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam, "Emotion and decision making," *Annu. Rev. Psychol.*, vol. 66, no. 1, pp. 799–823, Jan. 2015.

[110] G. Loewenstein and J. S. Lerner, "The role of affect in decision making," in *Handbook of Affective Sciences*, R. J. Davidson, K. R. Scherer, and H. H. Goldsmith, Eds. Oxford, England: Oxford University Press, 2003.

[111] L. F. Barrett, "The theory of constructed emotion: an active inference account of interoception and categorization," *Soc. Cogn. Affect. Neurosci.*, p. nsw154, Oct. 2016.

[112] R. W. Picard, *Affective Computing*, ser. The MIT Press. London, England: MIT Press, Jul. 2000.

[113] T. Singer and C. Lamm, "The social neuroscience of empathy," *Ann. N. Y. Acad. Sci.*, vol. 1156, no. 1, pp. 81–96, Mar. 2009.

[114] K. Crawford, *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, Apr. 2021.

[115] L. Floridi and J. W. Sanders, "On the morality of artificial agents," *Minds and Machines*, 2004.

[116] K. Hill, "A teen was suicidal. ChatGPT was the friend he confided in," *The New York Times*, Aug. 2025.

[117] T. Singleton, T. Gerken, and L. McMahon, "How a chatbot encouraged a man who wanted to kill the queen," *BBC News*, Oct. 2023.

[118] S. J. Paech, "Eq-bench: An emotional intelligence benchmark for large language models," 2023.

[119] F. Zhang, Z. Cheng, C. Deng, H. Li, Z. Lian, Q. Chen, H. Liu, W. Wang, Y.-F. Zhang, R. Zhang, Z. Guo, Z. Zhu, H. Wu, H. Wang, Y. Zheng, X. Peng, X. Wu, K. Wang, X. Li, J. Ye, and P.-A. Heng, "Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models," 2025.

[120] H. Hu, L. You, H. Xu, Q. Wang, F. R. Yu, F. Ma, Z. Cheng, Z. Lian, Y. Zhou, and L. Cui, "Emobench-m: Benchmarking emotional intelligence for multimodal large language models," 2025.

[121] E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, J. S. Prince, L. T. Dowdle, M. Nau, B. Caron, F. Pestilli, I. Charest, J. B. Hutchinson, T. Naselaris, and K. Kay, "A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence," *Nat. Neurosci.*, vol. 25, no. 1, pp. 116–126, Jan. 2022.

[122] S. E. J. de Vries, J. A. Lecoq, M. A. Buice, P. A. Groblewski, G. K. Ocker, M. Oliver, D. Feng, N. Cain, P. Ledochowitsch, and D. Millman, "A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex," *Nat. Neurosci.*, vol. 23, no. 1, pp. 138–151, 2020.

[123] C. Findling, F. Hubert, International Brain Laboratory, L. Acerbi, B. Benson, J. Benson, D. Birman, N. Bonacchi, E. K. Buchanan, S. Bruijns, M. Carandini, J. A. Catarino, G. A. Chapuis, A. K. Churchland, Y. Dan, F. Davatolhagh, E. E. J. DeWitt, T. A. Engel, M. Fabbri, M. A. Faulkner, I. R. Fiete, L. Freitas-Silva, B. Gercek, K. D. Harris, M. Häusser, S. B. Hofer, F. Hu, J. M. Huntenburg, A. Khanal, C. Krasniak, C. Langdon, C. A. Langfield, P. E. Latham, P. Y. P. Lau, Z. Mainen, G. T. Meijer, N. J. Miska, T. D. Mrsic-Flogel, J.-P. Noel, K. Nylund, A. Pan-Vazquez, L. Paninski, J. Pillow, C. Rossant, N. Roth, R. Schaeffer, M. Schartner, Y. Shi, K. Z. Socha, N. A. Steinmetz, K. Svoboda, C. Tessereau,

A. E. Urai, M. J. Wells, S. J. West, M. R. Whiteway, O. Winter, I. B. Witten, A. Zador, Y. Zhang, P. Dayan, and A. Pouget, "Brain-wide representations of prior information in mouse decision-making," *Nature*, vol. 645, no. 8079, pp. 192–200, Sep. 2025.

[124] E. M. Russek, D. Acosta-Kane, B. van Opheusden, M. G. Mattar, and T. L. Griffiths, "Time spent thinking in online chess reflects the value of computation," *Cogn. Sci.*, vol. 49, no. 10, p. e70119, Oct. 2025.

[125] H. J. Spiers, A. Coutrot, and M. Hornberger, "Explaining world-wide variation in navigation ability from millions of people: Citizen science project sea hero quest," *Top. Cogn. Sci.*, vol. 15, no. 1, pp. 120–138, Jan. 2023.

[126] P. Paquette, Y. Lu, S. Bocco, M. O. Smith, S. Ortiz-Gagne, J. K. Kummerfeld, S. Singh, J. Pineau, and A. Courville, "No press diplomacy: Modeling multi-agent gameplay," 2019. [Online]. Available: https://arxiv.org/abs/1909.02128

[127] J. E. Laird, A. Newell, and P. S. Rosenbloom, "Soar: An architecture for general intelligence," *Artificial Intelligence*, vol. 33, no. 1, pp. 1–64, 1987.

[128] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin, "An integrated theory of the mind," *Psychological Review*, vol. 111, no. 4, pp. 1036–1060, 2004.

[129] Y. LeCun, "A path towards autonomous machine intelligence," OpenReview, version 0.9.2, 2022, https://openreview.net/pdf?id=BZ5a1r-kVsf.

[130] A. Heim, "Yann LeCun's AMI Labs raises $1.03B to build world models," TechCrunch, Mar. 2026, https://techcrunch.com/2026/03/09/yann-lecuns-ami-labs-raises-1-03-billion-to-build-world-models/, accessed 2026-07-21.

[131] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," Sep. 2014.

[132] T. Nakai and S. Nishimoto, "Quantitative models reveal the organization of diverse cognitive functions in the brain," *Nat. Commun.*, vol. 11, no. 1, p. 1142, Mar. 2020.

[133] H. Jung, M. Amini, B. J. Hunt, E. I. Murphy, P. Sadil, Y. O. Halchenko, B. Petre, Z. Miao, P. A. Kragel, X. Han, M. O. Heilicher, M. Sun, O. G. Collins, M. A. Lindquist, and T. D. Wager, "Spacetop: A multimodal fMRI dataset unifying naturalistic processes with a rich array of experimental tasks," *Sci. Data*, vol. 12, no. 1, p. 1465, Aug. 2025.