

Cognitive Foundations for Reasoning and Their Manifestation in LLMs

Priyanka Kargupta¹◊, Shuyue Stella Li²◊, Haocheng Wang³,

Jinu Lee¹, Shan Chen⁴, Orevaoghene Ahia², Dean Light²,

Thomas L. Griffiths³, Max Kleiman-Weiner², Jiawei Han¹, Asli Celikyilmaz², Yulia Tsvetkov²

¹ University of Illinois Urbana-Champaign, ² University of Washington, ³ Princeton University, ⁴ Harvard University

◊ Equal contribution in alphabetical order.

Date: November 20, 2025

Correspondence: pk360@illinois.edu, stelli@cs.washington.edu

Code: <https://github.com/stellalisy/CognitiveFoundations>

Data: <https://huggingface.co/collections/stellalisy/cognitive-foundations>

Blogpost: <https://tinyurl.com/cognitive-foundations>

Abstract

Large language models solve complex problems yet fail on simpler variants, suggesting they achieve correct outputs through mechanisms fundamentally different from human reasoning. We synthesize cognitive science research into a taxonomy of 28 cognitive elements spanning computational constraints, meta-cognitive controls, knowledge representations, and transformation operations, then analyze their behavioral manifestations in reasoning traces. We propose a fine-grained cognitive evaluation framework and conduct the first large-scale analysis of 170K traces from 17 models across text, vision, and audio modalities, alongside 54 human think-aloud traces, which we make publicly available. Our analysis reveals systematic structural differences: humans employ hierarchical nesting and meta-cognitive monitoring while models rely on shallow forward chaining, with divergence most pronounced on ill-structured problems. Meta-analysis of 1,598 LLM reasoning papers reveals the research community concentrates on easily quantifiable behaviors (sequential organization: 55%, decomposition: 60%) while neglecting meta-cognitive controls (self-awareness: 16%, evaluation: 8%) that correlate with success. Models possess behavioral repertoires associated with success but fail to deploy them spontaneously. Leveraging these patterns, we develop test-time reasoning guidance that automatically scaffold successful structures, improving performance by up to 60% on complex problems. By bridging cognitive science and LLM research, we establish a foundation for developing models that reason through principled cognitive mechanisms rather than brittle spurious reasoning shortcuts or memorization, opening new directions for both improving model capabilities and testing theories of human cognition at scale.

Contents

1	Introduction	2
2	Formalizing Cognitive Foundations for Reasoning	4
2.1	Reasoning Invariants: Computational Constraints	6
2.2	Meta-Cognitive Controls: Executive Regulation	7
2.3	Reasoning Representations: Organizational Structures	8

2.4	Reasoning Operations: Transformation Procedures	10
3	Behavioral Manifestation in Humans and LLMs	11
3.1	Methodology	11
3.1.1	Data Collection	11
3.1.2	Fine-Grained Cognitive Element Annotation	12
3.1.3	Problem Type Classification & Response Evaluation	13
3.1.4	Cognitive Structure Extraction	13
3.2	Experimental Setup	14
3.2.1	Dataset Composition	14
3.2.2	Analysis Dimensions	15
3.3	Results & Analyses	15
3.3.1	Distribution of Cognitive Elements	15
3.3.2	Behavioral Structures	17
3.4	Comparison with Humans	18
4	Eliciting Cognitive Reasoning Structures	19
5	Cognitive Element Considerations in LLM Research Design	21
6	Opportunities and Challenges	22
A	Appendix	35

1 Introduction

Humans are capable of extrapolating from their existing knowledge to unfamiliar scenarios and generating new knowledge—a process that constitutes reasoning (Lombrozo, 2024). In contrast, large language models (LLMs) exhibit a form of cognitive dissonance unprecedented in human cognition: they master complex skills (Chervonyi et al., 2025; Jiang et al., 2024) while lacking simpler prerequisite ones (McCoy et al., 2024; Dziri et al., 2023; Berglund et al., 2023), and solve challenging problems while failing on trivial variants (Shao et al., 2025; Li et al., 2025b). This dissociation between high benchmark performance and the lack of generalization suggests that models may be achieving correct outputs through mechanisms fundamentally different from the robust cognitive structures underlying human reasoning. Current training and evaluation paradigms reward reasoning outcomes without examining the cognitive processes that produce them (Lambert et al., 2024), and therefore cannot distinguish between genuine reasoning and memorization (Wu et al., 2025). This creates a measurement crisis: we lack the conceptual framework to characterize what reasoning behaviors should manifest in models and the empirical methods to assess whether they do.

To understand how cognitive elements manifest in reasoning behaviors, consider a child playing with LEGO blocks. Figure 1 illustrates how the child orchestrates multiple cognitive elements: establishing goals, decomposing the task into parts, verifying connections, monitoring progress, and reformulating failed approaches. This flexible coordination of diverse cognitive processes characterizes reasoning. However, contemporary research on LLM reasoning only attempts scattered investigations of specific behaviors such as verification (Gandhi et al., 2025), decomposition (Xu et al., 2025), and self-monitoring (Marjanović et al., 2025). Our meta-analysis of 1,598 arXiv LLM reasoning papers on LLM reasoning reveals this fragmentation quantita-

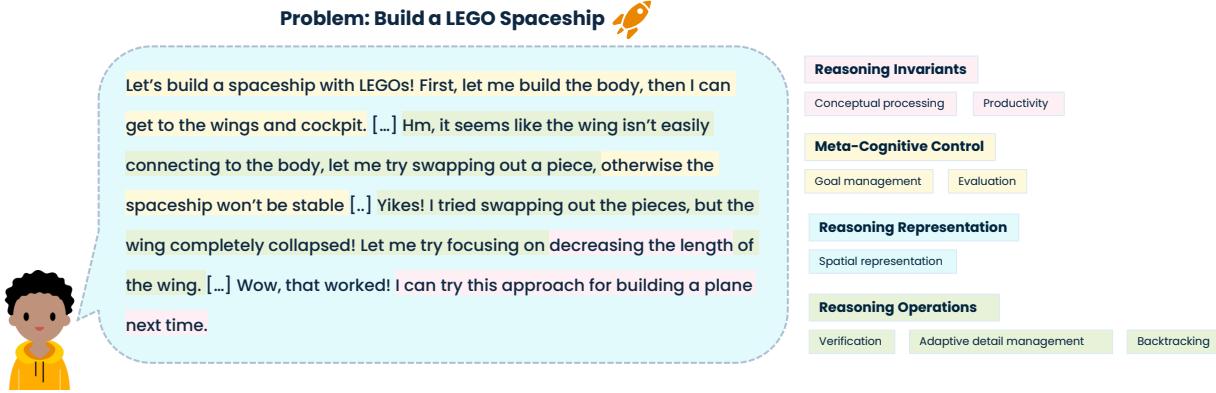


Figure 1: An example of reasoning behaviors present in a reasoning trace for building a LEGO spaceship. We organize the different behaviors along the four dimensions of our taxonomy, as shown in Table 1.

tively (Figure 2): 55% examine sequential organization (e.g. reasoning step-by-step), while only 8% study backward chaining (e.g. reverse engineering the solution from the answer). This concentration on easily quantifiable behaviors creates a striking asymmetry when compared with human reasoning research, which encompasses diverse phenomena, including analogical transfer (Gentner, 1983; Holyoak & Thagard, 1989), causal inference (Sloman & Sloman, 2009; Gopnik et al., 2004), representational flexibility (Ohlsson, 1992; Knoblich et al., 1999), and meta-cognitive control (Nelson, 1990; Flavell, 1979). Each of these captures different facets of how humans navigate uncertainty, integrate knowledge, and adapt to novel situations. Without a comprehensive framework spanning this range of cognitive elements, we risk optimizing what we measure rather than what matters, potentially mistaking narrow competence for broad reasoning capability.

To address this gap, we **introduce a unified taxonomy of cognitive foundations** by synthesizing established theories of problem-solving, mental representation, and meta-cognition (Sweller, 1988; Johnson-Laird, 1983; Fodor, 1975). The taxonomy includes four dimensions: **1) Reasoning invariants** capture fundamental properties that must hold for valid reasoning, including logical coherence (not simultaneously believing "the design is stable" and "the design will collapse"), compositionality (understanding "red LEGO cockpit" by combining concepts of color, material, and function). **2) Meta-cognitive controls** encompass the executive functions that select and monitor reasoning strategies, such as recognizing when you lack necessary pieces (self-awareness), deciding whether to plan the entire design upfront versus building exploratively (strategy selection). **3) Reasoning representations** describe how knowledge is organized: hierarchically (decomposing "spaceship" into "body," "wings," "cockpit"), causally (understanding that inadequate support causes structural collapse), or spatially (tracking how pieces connect in 3D space). **4) Reasoning operations** specify the procedures that construct and transform these representations, such as verifying each connection works, backtracking when a wing design fails, or abstracting the principle that heavier components need more support points. Organizing these theories through Marr's (1982) levels of analysis (focusing on computational and algorithmic levels) yields 28 specific cognitive elements spanning the space of human

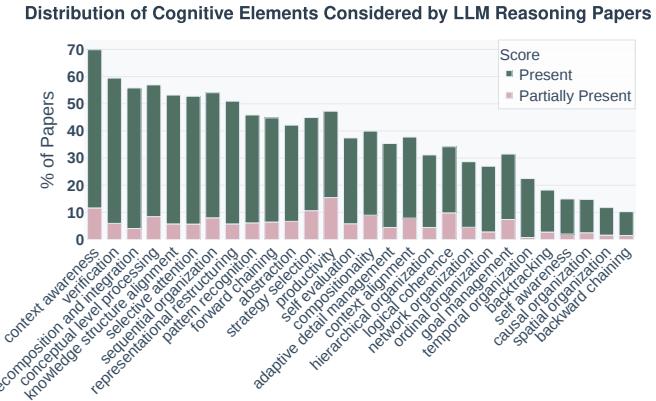


Figure 2: Distribution of cognitive element presence across 1,598 arXiv LLM Reasoning papers. *Partially present* indicates that there is evidence that the behavior was considered in the design (motivation, method, evaluation) of the paper but was not the primary focus. *Present* indicates that there is evidence that the behavior was a conscious and significant design decision. Details provided in Section 5.

reasoning capabilities (Table 1). This taxonomy represents a theoretical contribution: we synthesize existing cognitive science literature into an analytical vocabulary for studying machine reasoning, providing a bridge between human cognitive research and LLM evaluation.

With this taxonomy, we conduct the **first large-scale empirical comparison of cognitive elements in human versus LLM reasoning** across diverse problem types (Jonassen, 2000). We analyze 170K reasoning traces from 17 models across text, vision, and audio modalities, complemented by 54 human think-aloud reasoning traces (Table 2). Using fine-grained span-level annotation validated by human evaluation, we identify which of the 28 cognitive elements appear in each trace, where they occur, and how they are sequenced in thought. Our analysis reveals fundamental differences in reasoning presence and structures. We apply our analysis framework to distinguish between elements that are frequently exhibited by models and those that are most strongly correlated with correct outcomes. Through our analysis, we reveal that **models consistently employ behaviors which are not the most conducive to success**. We specifically observe that as problems become more ill-defined and non-verifiable (e.g., open-ended dilemmas, diagnosing and solving multifaceted problems), models narrow their selection of elements to rigid strategies despite *a broader, more diverse usage of cognitive elements being empirically shown as a more effective strategy*. Furthermore, we devise a novel structural representation to encode the structure of cognitive elements within reasoning traces and find that models frequently choose reasoning structures which differ from successful structure. Finally, we apply this same analysis framework to comparing humans and LLMs. We observe that humans employ more frequent abstract and conceptual-level processing, while LLMs rely more heavily on shallow sequential forward chaining with limited corrective structures based on problem type and structure.

Leveraging these behavioral correlations, we introduce **test-time reasoning guidance** as a targeted intervention to improve model performance by explicitly scaffolding cognitive patterns predictive of reasoning success. For instance, on dilemma problems where successful DeepSeek-R1 traces exhibit the behavioral sequence of self-awareness followed by hierarchical representation construction and then problem decomposition, we prompt the model to follow this structure. Evaluation demonstrates that test-time reasoning guidance improves performance by up to 66.7% on ill-structured problems while maintaining baseline performance on well-structured ones. This improvement confirms that models possess reasoning capabilities not spontaneously expressed, and that understanding cognitive behavioral patterns can inform more effective model interaction strategies.

This work establishes a taxonomy of cognitive foundations that bridges human reasoning research and LLM evaluation, providing a unified vocabulary for characterizing reasoning processes beyond performance metrics. By synthesizing cognitive science theories through Marr’s levels of analysis, we identify 28 reasoning behaviors spanning computational constraints, meta-cognitive controls, knowledge representations, and transformation operations. We conduct the first large-scale empirical analysis across 170K model and human reasoning traces¹, revealing systematic structural differences: humans employ hierarchical nesting and meta-cognitive monitoring while models rely on shallow sequential forward chaining, with divergence most pronounced on ill-structured problems. Leveraging these patterns, we introduce cognitively-grounded reasoning guidance to the model that automatically scaffolds successful reasoning structures, improving performance by up to 60% on complex problems and confirming that models possess latent capabilities not spontaneously expressed. Our analysis of 1,598 LLM reasoning papers reveals critical between research emphasis and behavioral requirements. This framework enables researchers to ask targeted questions about LLM reasoning: which behaviors emerge from scale and which require architectural innovation, do different training paradigms produce different cognitive profiles, and how do reasoning structures transfer across domains and modalities? These questions become answerable when we have a shared vocabulary for characterizing reasoning processes. Our taxonomy and annotation methodology provide that vocabulary, grounded in decades of cognitive science research and validated through large-scale empirical analysis of both human and machine reasoning traces.

2 Formalizing Cognitive Foundations for Reasoning

Understanding the extent to which language models reason requires a theory of what reasoning is. Without principled criteria distinguishing reasoning from pattern matching, evaluation isn’t possible. Decades of

¹Publicly released at <https://huggingface.co/collections/stellalisy/cognitive-foundations>

Table 1: Taxonomy of cognitive reasoning behaviors, organized along two main dimensions: Cognitive Properties (stable internal constraints) and Cognitive Capabilities (deployable context-sensitive abilities).

A. Reasoning Invariants: "Always-true" constraints or quality criteria the system maintains across reasoning steps.		
Logical coherence	Maintain consistency across reasoning steps and contexts (Fodor & Pylyshyn, 1988).	
Compositionality	Build complex ideas from simpler components (Fodor, 1975).	
Productivity	Formulate an indefinite number of thoughts or solutions using a finite set of elements (Halford, 1989).	
Conceptual processing	Operating over abstract representations before linguistic expression (Halford, 1989).	
B. Meta-Cognitive Controls: Higher-order abilities that select, monitor, and adapt processes.		
Self-awareness	Assess own knowledge state, capabilities, and task solvability (Wicklund, 1979).	
Context awareness	Perceive, understand, and navigate one's circumstances (including other agents) (Frith & Frith, 2007).	
Strategy selection	Choose & explore reasoning approaches suited to task and domain demands (Lieder & Griffiths, 2017).	
Goal management	Establish, maintain, and adjust goals throughout the reasoning process (Griffiths et al., 2019).	
Evaluation	Assess & adapt to the quality, efficiency, and progress of one's reasoning (Fleming & Daw, 2017).	
C. Reasoning Representations: The formats and organizational patterns used to encode and relate knowledge and steps.		
Structural Organization	Sequential organization Hierarchical organization Network organization	Order steps where sequence matters (Skinner, 1953). Nest concepts in parent-child relationships (Galanter et al., 1960). Link concepts through multiple relationship types (Quillan, 1966).
Conceptual Organization	Ordinal organization Causal organization Temporal organization Spatial organization	Arrange elements by relative order or rank (Stevens, 1946). Connect elements through cause-effect relations (Heider, 1958). Order elements by before-after relations (Ebbinghaus, 1885). Structure elements by spatial relationships (Tolman, 1948).
D. Reasoning Operations: Goal-directed actions that construct, evaluate, modify, and navigate reasoning representations.		
Representation Selection	Context alignment Knowledge alignment	Align to task and situational demands (Gick & Holyoak, 1980). Align to domain-specific structures & relations (Chi et al., 1981).
Representation Evaluation	Verification	Check reasoning steps against pre-determined criteria (Flavell, 1979).
Representation Modification	Selective attention Adaptive detail mgmt. Decomposition and integration Representational restructuring Pattern recognition Abstraction	Focus on relevant details and filtering noise (Broadbent, 1958). Adjust granularity based on task demands (Rosch, 1978). Divide problems and synthesizing subsolutions (Newell et al., 1959). Reformulate problems for new insights (Wertheimer, 1945). Detect recurring structures across contexts (Selfridge, 1959). Generalize from specific cases (Hull, 1920).
Representation Navigation	Forward chaining Backward chaining Backtracking	Reason from known facts toward goals (Huys et al., 2012). Work backward from goals to prerequisites (Park et al., 2017). Revisit and correcting prior reasoning paths (Nilsson, 1971).

research in Cognitive Science on problem-solving, mental representation, and metacognition provide a powerful starting point. The cognitive revolution of the 1950s introduced a powerful framework: the mind as an information-processing system. Fodor's (1975) Language of Thought Hypothesis proposed that thinking is a computational process operating over compositional representations. Complex thoughts comprise simpler components combined through systematic rules, enabling humans to generate infinitely many novel thoughts from finite conceptual primitives. We are not "hapless tourists bound by a limited phrasebook of ideas" (Frankland & Greene, 2020), but fluent speakers of an internal, compositional language. However, this view of the mind as a formal logic engine could not explain empirical findings that emerged from psychology laboratories. Wason (1968) demonstrated a "content effect", where people systematically fail to apply logical rules in an abstract setting but succeed when the identical task is presented as a familiar social rule. Tversky & Kahneman (1974) showed that humans systematically use heuristics and mental shortcuts that violate basic principles of logic and probability theory (Ragni et al., 2017). These results have led to many alternative theoretical accounts. Mental Models Theory proposed that reasoning constructs and manipulates semantic simulations of the world rather than applying syntactic inference rules (Johnson-Laird, 1983; 2010). Dual-Process Theory distinguished fast, intuitive processing from slow, deliberative reasoning (Evans, 2003; Evans & Stanovich, 2013). Bayesian approaches reframed rationality as probabilistic inference under uncertainty,

arguing that reasoning updates probabilistic beliefs based on new, ambiguous, and noisy data in ways that approximate Bayesian inference (Griffiths et al., 2024; Jacobs & Kruschke, 2011; Oaksford & Chater, 2009).

This theoretical diversity in cognitive science, while valuable, has contributed to fragmentation in LLM reasoning research. Gandhi et al. (2025) characterizes reasoning through verification, subgoal setting, backtracking, and backward chaining, while Lee et al. (2024) focuses on logical coherence, compositionality, and productivity. Models appear capable under one framework yet fail under another, and researchers lack a shared vocabulary for diagnosing where and why reasoning breaks down. We address this by organizing cognitive theories using Marr’s (1982) levels of analysis and proposing a unified framework for cognitive elements. The *computational level* defines the goal of the system: what is being computed and why. The *algorithmic and representational level* specifies the process: what representations are used for the input/output and what algorithms transform these representations to achieve the computational goals. The *implementation level* concerns the physical realization about how representations and algorithms are instantiated in neural hardware or silicon. For our purposes, we focus on the first two levels. At the computational level, reasoning must satisfy certain fundamental properties to maintain consistency, resolve contradictions, and combine elements compositionally. At the algorithmic level, these goals are realized through specific knowledge structures and the processes that manipulate them (Peebles & Cooper, 2015; Krafft & Griffiths, 2018).

Our proposed framework for cognitive foundations consists of four dimensions (detailed in Table 1). 1. *Reasoning invariants* specify computational goals—properties that must hold for valid reasoning. 2. *Meta-cognitive controls* select and monitor processes, determining which strategies to deploy and when to adapt. 3. *Reasoning representations* describe how knowledge is organized. 4. *Reasoning operations* specify procedures that construct, evaluate, and transform representations. These dimensions span 28 cognitive elements, providing vocabulary for diagnosing which aspects of human reasoning manifest in language models.

A familiar scenario illustrates how these dimensions interact: a child playing with a pile of LEGO® bricks and building something novel, like a "spaceship" (Figure 1). This creative process is an example of the reasoning invariants of compositionality and productivity, where the goal is to generate a complex, new structure (the spaceship) by combining a finite set of simple components (a few different types of toy bricks). This computational goal is managed by meta-reasoning controls. The child must first decide on a strategy: "Should I have a rough plan first, or go directly into building?" "Should I start with the main body or the wings?" Throughout the process, they must constantly monitor their progress and evaluate their spaceship. For instance, they might observe that the wings don't look sturdy enough, leading them to reinforce the structure with more blocks. This entire cycle of planning, building, and evaluation relies on the child forming and manipulating a dynamic 3D mental model of the target spaceship, especially the spatial relationships between components. This mental model shifts from a vague concept to a concrete plan through a series of operations. For instance, the child would try out different combinations of pieces, constantly assembling, testing, and then backtracking to refine their final design.

This LEGO example illustrates that reasoning arises from the dynamic interplay of all four dimensions, because the child's compositional goal requires strategic choices about how to build, which determine what representations to construct, which are then transformed through coordinated operations. Current research tends to isolate individual behaviors—studying verification independently from representation selection, or decomposition separately from meta-cognitive monitoring. However, effective reasoning requires these dimensions to work in concert. The computational drives and constraints from invariants are managed by meta-cognitive controls, which in turn direct the algorithmic-level shifts in representation and the operations that enact them. No single dimension suffices. In the following subsections, we will unpack each dimension in detail, drawing on research in cognitive science to explain what each dimension contributes and examining how each one manifests in current language models.

2.1 Reasoning Invariants: Computational Constraints

At Marr’s computational level, we investigate the fundamental constraints that an ideal solution to any reasoning problem must satisfy. Fodor’s (1975) language of thought hypothesis identifies these constraints: valid reasoning manipulates structured representations according to compositional rules, producing thoughts that are logically coherent (free from contradiction), compositional (complex meanings built from simpler

parts), and productive (capable of generating unbounded novel inferences from finite primitives). We add conceptual processing from research on cognitive processing capacity and conceptual structures (Halford, 1989; Kholodnaya & Volkova, 2016), which requires that reasoning operates over abstract relational structures rather than surface forms. These four invariants specify the computational constraints that define valid reasoning solutions.

Logical coherence maintains consistency across reasoning steps and contexts (Rips, 1983; Thagard, 2002). In the LEGO example, if the child simultaneously believes "this wing design is stable" and "this wing design will collapse," reasoning breaks down. The contradiction creates cognitive dissonance that must be resolved by revising the stability judgment, reinforcing the structure, or reconsidering the design (Festinger, 1962; Harmon-Jones & Mills, 2019). This pressure to restore consistency shapes how reasoners revise mental models and update inferences. Humans cannot comfortably hold contradictory beliefs simultaneously; the drive to resolve inconsistencies is a defining feature of reasoning.

Compositionality enables building complex ideas from simpler components through rule-governed combination (Fodor, 1975; 2001; Fodor & Pylyshyn, 1988). The LEGO spaceship emerges from systematically combining individual bricks. The structure of the whole derives predictably from the properties of its parts and how they connect. Understanding "red cockpit with transparent dome" requires combining concepts of color, component type, and material in ways that preserve each element's meaning. The meaning of a complex expression derives from the meanings of its constituents and their mode of combination (Russin et al., 2024). This property drastically increases the learning efficiency and expressive power of the cognitive system, allowing finite mental resources to produce unbounded conceptual richness (Frankland & Greene, 2020; Piantadosi, 2011).

Productivity extends compositionality by enabling the generation of infinitely many novel thoughts from finite primitives (Fodor, 1975). This generative power is intrinsically linked to systematicity, the capacity to entertain a set of thoughts given the ability to entertain related ones (Fodor & Pylyshyn, 1988). The recursive nature of human cognition provides the mechanism necessary for this unlimited generative power (Hauser et al., 2002). Having understood how to build a spaceship, the child can apply the same structural principles to generate a plane, a castle, or a dinosaur. This capacity to *produce* and understand unbounded thoughts from finite components is central to human reasoning. Once we grasp the compositional structure for combining LEGO blocks to achieve a functional goal, we spontaneously recognize how to apply that structure to new contexts without requiring explicit instruction for each instance. Productivity distinguishes genuine understanding from memorized responses.

Conceptual processing operates over abstract semantic relations rather than surface forms (Halford, 1989; Kholodnaya & Volkova, 2016; Baron, 2008). Having this conceptual depth enables generalization beyond shallow pattern matching of concrete, low-level objects. In the LEGO example, the child is not just combining bricks together based on their colors or shapes. Instead, they are operating on the abstract concept of a spaceship. This concept brings with it a set of functional properties that guide the building process: a spaceship needs a cockpit, a body, and wings.

These four invariants are interconnected requirements rather than independent properties. Coherence constrains which compositional structures are valid. Productivity depends on compositional structure to generate novel thoughts systematically. Conceptual processing ensures reasoning operates over meaningful relationships rather than surface patterns. Together, these invariants define the computational constraints that valid reasoning must satisfy. However, invariants specify what reasoning must achieve without addressing the mechanisms that recognize when constraints are violated, select appropriate strategies, or monitor progress toward goals. These regulatory mechanisms constitute meta-cognitive controls.

2.2 Meta-Cognitive Controls: Executive Regulation

Meta-cognitive controls constitute the executive functions that select, monitor, and adapt reasoning processes (Fleming, 2024). While invariants define validity criteria, controls orchestrate the reasoning process itself.

Self-awareness stands as the foundation: the capacity to assess one's own knowledge state, capabilities, and the solvability of a task (Neisser, 1988; Wicklund, 1979; Leary & Buttermore, 2003). Rochat (2003)

describes this as "arguably the most fundamental issue in psychology, from both a developmental and evolutionary perspective". In the LEGO example, the child engages in internal assessment: "Am I good at building spaceships?" or "Do I know what a spaceship looks like?" This metacognitive evaluation enables strategic deployment of cognitive resources.

Context awareness perceives and responds to situational demands, environmental constraints, and the presence of other agents (Boyd et al., 2011; Frith & Frith, 2007; Lei, 2023). In the LEGO example, context fundamentally shapes the task. Building alone for fun permits flexible exploration. Building with a friend introduces social dynamics requiring cooperation and negotiation. Building in a timed contest with limited bricks demands prioritizing speed and efficiency over creative exploration. Context awareness determines which strategies are appropriate and which goals are worth pursuing (Milli et al., 2021).

Given this awareness of self and context, the child engages in **strategy selection**, choosing an approach suited to task demands (Lieder & Griffiths, 2017; 2020; Mata et al., 2011). They might adopt an exploratory, bottom-up strategy, combining bricks and allowing the design to emerge organically. Alternatively, they could choose a planned, top-down strategy, first visualizing the spaceship mentally and then systematically finding bricks to realize each component. With limited bricks, they might adopt a resource-first strategy, sorting pieces to assess availability before committing to a design. With time constraints, they might use a schema-based strategy, quickly replicating a remembered design.

These strategic choices are accompanied by **goal management**: the process of setting, sequencing, and dynamically adjusting goals throughout reasoning (Griffiths et al., 2019; Dolan & Dayan, 2013; Cushman & Morris, 2015). The abstract goal "build a spaceship" decomposes into manageable sub-goals: construct the main body, add wings, build the cockpit, and insert pilots. Goal management is inherently dynamic, corresponding to strategy shifts due to changes in context and self-awareness. While working on wings, the child might spot a clear brick perfect for a windshield, pausing the current task to insert a new sub-goal. Conversely, if attaching wheels proves too difficult, the child abandons this sub-goal to prioritize completing the rest of the ship.

Progress toward these shifting goals is monitored through **evaluation**: assessing the quality, efficiency, and coherence of the emerging solution (Yeung & Summerfield, 2012; Fleming & Daw, 2017; Stipek et al., 1992). The child continuously evaluates: Does the wing design look stable? Is this approach too slow? Should I try a different configuration? This ongoing assessment determines whether to persist with the current approach or adapt strategies and goals.

These five controls form an integrated executive system operating in coordinated fashion. Self-awareness detects capabilities and limitations. Context awareness identifies situational demands. Strategy selection responds by choosing appropriate approaches. Goal management directs the response through structured sub-goals. Evaluation monitors progress and triggers adaptation when needed. In the LEGO example, these controls work together: the child assesses their skill level (self-awareness), recognizes time and resource constraints (context awareness), selects an appropriate strategy (strategy selection), breaks the task into sub-goals (goal management), and monitors whether the emerging structure matches their mental model (evaluation). Yet controls alone do not constitute reasoning. They govern processes but do not specify the representational structures over which those processes operate. A reasoner must organize knowledge in some form, and the choice of structure profoundly shapes reasoning effectiveness.

2.3 Reasoning Representations: Organizational Structures

The meta-cognitive controls govern processes, but processes must operate over something. Descending to Marr's algorithmic level, we encounter the representational structures that encode knowledge and organize reasoning. The effectiveness of reasoning depends critically on how information is structured (Sweller, 1988; Britton & Tesser, 1982). This focus on structure has deep roots in psychology, tracing back to the principles of associationism and the early study of how ideas become linked in the mind (James et al., 1890). Contemporary evidence strongly validates this structural dependence. Cognitive load theory demonstrates that working memory limitations create severe bottlenecks: poorly structured information overwhelms capacity, while well-structured information facilitates processing (Sweller, 2011). Various representation structures have been proposed. Production systems represent procedural knowledge via a set of If-Then rules. They

form the basis of early cognitive modeling systems like ACT-R (Newell & Simon, 1972). Prototype theory (Rosch, 1975; Rosch & Mervis, 1975) represents categories by their most typical member, while exemplar theories (Medin & Schaffer, 1978) represent a collection of all remembered instances of that category. Semantic network theories model human knowledge as nodes connected by typed relations, capturing both hierarchical and associative organization (Quillian, 1966; Collins & Loftus, 1975; Steyvers & Tenenbaum, 2005). More generalized structures, such as frames and schemas, encode stereotyped knowledge about common situations, guiding prediction and inference by providing default expectations (Minsky, 1974; Bartlett, 1932). Mental models research shows that humans construct internal representations reflecting structural and causal relationships in the domains they reason about (Johnson-Laird, 1983). Here, we organize representations along two dimensions: structural organization concerns how elements connect, while conceptual organization concerns how meaning is arranged.

Structural organization specifies the architecture through which elements relate. The simplest form is **sequential organization**, which orders steps where sequence matters (Skinner, 1953). The procedure for starting a car, the temporal flow of historical events, and the steps in a recipe all depend on maintaining proper order. But many reasoning tasks demand richer connectivity (Lashley et al., 1951; Rosenbaum et al., 2007). **Hierarchical organization** nests concepts in parent-child relationships, enabling decomposition of complex wholes into manageable parts (Galanter et al., 1960; Botvinick et al., 2009; Haupt, 2018). Biological taxonomies classify organisms into kingdom, phylum, class, and order; problem-solving decomposes complex tasks into manageable subtasks; and administrative structures organize authority and responsibility through nested levels. This parent-child structure proves so cognitively natural that it appears across virtually every domain of human knowledge organization. Yet even hierarchies have limitations. Many domains resist strict tree structures because elements relate through multiple relationship types simultaneously. **Network organization** captures this richer connectivity (Quillian, 1966; Shafto et al., 2008). In understanding an ecosystem, organisms connect through predation, competition, symbiosis, and energy flow. No single hierarchy can represent all these relations; only a network preserves the full relational structure. In the LEGO example, the child's mental model exhibits hierarchical decomposition of the spaceship into body, wings, and cockpit, while simultaneously maintaining network relationships that specify how components physically connect and structurally support each other.

Conceptual organization structures meaning rather than architecture, specifying the semantic relationships that give representations their inferential power. When the child reasons about alternative wing designs for their LEGO spaceship, **ordinal organization** allows ranking them from best to worst, most stable to least stable (Stevens, 1946). But ranking alone provides no explanatory depth. **Causal organization** connects elements through cause-effect relations, revealing why one design outperforms another (Heider, 1958; Khemlani et al., 2014). Inadequate wing reinforcement causes structural instability, which prevents stable flight. This causal chain guides both diagnosis (why did it fail?) and intervention (how to fix it?). Causal reasoning interacts intimately with **temporal organization**, which orders events by before-after relations (Ebbinghaus, 1885; Bartlett, 1932; Hoerl & McCormack, 2019). The child must attach the wings before adding the cockpit and test stability before declaring the design complete. Temporal constraints shape planning, while causal understanding determines which temporal orderings make sense. Finally, **spatial organization** structures elements geometrically, capturing relationships like adjacency, containment, and orientation (Tolman, 1948; Shepard & Metzler, 1971; Landau & Jackendoff, 1993). The child reasons about which pieces fit where, how rotating a component changes available attachment points, where weight distribution affects balance (Shelton et al., 2022; Cortesa et al., 2017). In practice, the LEGO example reveals how these conceptual organizations interweave: spatial reasoning determines physical fit, causal reasoning predicts structural stability, and temporal reasoning sequences construction steps. The child fluidly transitions between organizational schemes as task demands shift.

These structures are not static containers but dynamic scaffolds that reasoning actively constructs and transforms. Reasoning does not simply retrieve and display pre-formed representations. It constructs, evaluates, modifies, and navigates them through cognitive operations.

2.4 Reasoning Operations: Transformation Procedures

Representations provide the scaffolds; operations are the processes that manipulate them. At the algorithmic level, reasoning unfolds through goal-directed procedures that construct new representations, evaluate existing ones, modify them adaptively, and navigate through them strategically (McClelland et al., 2010; Johnson-Laird, 1983). These operations are enacted under meta-cognitive control and constrained by reasoning invariants. While humans possess a large repertoire of reasoning operations, we focus on the four most fundamental clusters that characterize reasoning across domains: selection, evaluation, modification, and navigation.

Representation selection operations align structures to task and domain demands, a process rarely conscious but profoundly shaping subsequent reasoning. **Context alignment** chooses organizational schemas fitting the problem (Gick & Holyoak, 1980; 1983; Gentner, 1983): temporal for historical explanation, causal for scientific reasoning, spatial for navigation. **Knowledge alignment** maps problems onto domain-specific schemas (Chase & Simon, 1973; Chi et al., 1981; Schoenfeld, 2014). A physician diagnosing symptoms activates medical taxonomies; an auto mechanic activates mechanical systems and failure modes. In the LEGO example, the child implicitly selects hierarchical decomposition combined with spatial organization as appropriate for the construction task, a choice that determines which reasoning moves become natural.

Representation evaluation verifies reasoning steps against predetermined criteria. **Verification** checks intermediate inferences for consistency, plausibility, and coherence with known facts (Flavell, 1979; Goldstein et al., 2010). The child might verify that two pieces actually connect as expected by testing the attachment, or recount the number of attachment points to confirm stability. Humans continuously verify their reasoning, though the sophistication and systematicity vary with expertise (Simon & Simon, 1978). Novices may fail to catch errors that experts immediately flag.

Representation modification operations adaptively change representational form. **Selective attention** focuses on relevant features while filtering noise. In the LEGO example, the child can attend to wing structure and stability while ignoring piece color or aesthetic details. This ability to focus resources and filter irrelevant input is fundamental to information processing, established by classic models demonstrating a bottleneck in early sensory processing that requires a selection mechanism (Broadbent, 1958; Treisman, 1964). **Adaptive detail management** adjusts granularity based on task demands (Rosch, 1978). Individuals dynamically shift between fine-grained and global representations based on cognitive goals, zooming in to examine how specific connection points distribute weight or zooming out to assess overall balance. **Decomposition and integration** break problems into subproblems and synthesize solutions (Newell et al., 1959). Decomposition is a core mechanism of problem-solving, initially formalized through means-ends analysis and the creation of subgoals to manage task complexity. Integration represents the essential process of synthesizing findings across these subgoals to form a complete solution. **Representational restructuring** reframes problems for new insight, fundamentally shifting how the problem is conceptualized and often marking breakthrough moments in reasoning (Wertheimer, 1945; Köhler, 1925; Braun et al., 2010). **Pattern recognition** detects recurring structures, enabling reuse of solution templates (Selfridge, 1959; Posner & Keele, 1968). This operation is the foundation of expert knowledge, enabling fast and accurate classification of novel stimuli by matching them against stored prototypes or templates. **Abstraction** generalizes from specific instances, distilling common, invariant features from a series of concrete examples and deriving principles from a set of actions or operations (Hull, 1920; Piaget, 1952). After several construction attempts, the child may abstract the principle that heavier components require proportionally more connection points to remain stable.

Representation navigation includes operations that traverse knowledge and inference structures. **Forward chaining** reasons from known facts toward goals: "These pieces connect this way; therefore this structure is possible" (Huys et al., 2012). This operation is a key component of early production systems and expert system architectures, where the system begins with facts and applies rules iteratively until a conclusion is reached or a goal state is satisfied (Newell & Simon, 1972; Davis & King, 1984) **Backward chaining** works from goals to prerequisites: "To make wings stable, I need reinforcement; what pieces provide that?" (Park et al., 2017). This goal-driven approach is highly efficient in situations where the number of possible outcomes is large, but the goal is clearly defined (Charniak, 1985; Newell et al., 1959). **Backtracking**

revisits earlier reasoning upon detecting errors (Qin et al., 2025). This crucial cognitive control mechanism often relies on depth-first search, where, upon hitting a dead-end, the system returns to the most recent decision point to explore alternatives (Nilsson, 1971). In the LEGO example, the child uses forward chaining to build, backward chaining to plan, and backtracking when a design fails. These navigation operations determine the path through problem space.

These operations do not occur in isolation. In the LEGO example, representation selection chooses hierarchical decomposition, forward chaining builds the body, verification checks that pieces connect properly, backtracking returns to an earlier design choice when the structure wobbles, restructuring introduces new design ideas, pattern recognition suggests a previously worked design, and abstraction extracts a principle for future builds. The operations form an integrated toolkit, flexibly deployed under meta-cognitive control to manipulate representations while satisfying invariants. Together with the representational structures from Section 2.3, they constitute Marr’s algorithmic level. Table 1 presents our complete taxonomy: invariants define computational constraints, controls select and monitor processes, representations encode knowledge structures, and operations transform those structures. This framework provides the analytical vocabulary for examining how these 28 reasoning behaviors manifest in human versus LLM reasoning. Section 3 presents our investigation of behavioral presence, patterns, and their relationship to reasoning success across diverse problem types.

3 Behavioral Manifestation in Humans and LLMs

3.1 Methodology

3.1.1 Data Collection

Table 2: Overview of Data Sources and Models by Modality

Modality	Dataset	#	Model	#	# Traces
Text	GeneralThought (Taylor, 2024)	10,322	Qwen3 (8B, 14B, 32B) (Yang et al., 2025)	3	31,836
	ClaimSpect (Kargupta et al., 2025)	290	DeepSeek-R1-Distill-Qwen2.5 (1.5B, 7B, 14B, 32B) (DeepSeek-AI et al., 2025)	4	42,448
			DeepSeek-R1-Distill-Llama3 (8B, 70B) (DeepSeek-AI et al., 2025)	2	21,224
			DeepScaleR-1.5B-Preview (Luo et al., 2025)	1	10,612
			s1.1-32B (Muennighoff et al., 2025)	1	10,612
			OpenThinker-32B (Guha et al., 2025)	1	10,612
			DeepHermes-3-Llama-3-8B-Preview (Teknium et al., 2025)	1	10,612
			DeepSeek-R1 (DeepSeek-AI et al., 2025)	1	10,612
	Total (Text Data)	10,612	Total (Text Models)	14	148,568
	BLAB (Ahia et al., 2025)	417	Qwen/Qwen3-Omni-30B-A3B-Thinking Yang et al. (2025)	1	4,917
Audio	MMAR (Ma et al., 2025)	888			
	MMAU-Pro (Kumar et al., 2025)	3,612			
	Total (Audio Data)	4,917	Total (Audio Models)	1	4,917
Image	Zebra-CoT (Li et al., 2025a)	18,000	GPT 4.1 and Gemini 2.5 pro (Only used for refinement)	2	18,000
	Total (Image Data)	18,000	Total (Image Models)	2	18,000
	Grand Total (All Data)	33,529	Grand Total (All Models)	17	171,485

LLM Reasoning Data. We analyze 14 open-weight **text reasoning models** spanning multiple architecture families and training paradigms. These include: Qwen3 hybrid models with thinking mode, R1-Distilled models built on Qwen 2.5, R1-Distilled models built on Llama 3, several community-developed reasoning-tuned models (DeepScaleR-1.5B-Preview, s1.1-32B, OpenThinker-32B, DeepHermes-3-Llama-3-8B-Preview), and the frontier DeepSeek-R1 (671B), which also serves as the teacher model for R1-Distilled variants. **Text-only reasoning problems** consist of 10,612 questions sampled from the **GeneralThought** (Taylor, 2024) and **ClaimSpect** (Kargupta et al., 2025) datasets. The **task** label in the dataset is used

to down-sample extremely common tasks (e.g., arithmetic, simple logic) to maintain a balanced representation. We use released DeepSeek-R1 traces from GeneralThought due to compute constraints. While GeneralThought provides a wide variety of tasks and domains, the dataset primarily focuses on verifiable tasks and consequently lacks problems of the type Dilemma². Thus, we convert 290 real-world biomedical and geopolitical "nuanced" claims from ClaimSpect into questions to supplement the text-reasoning category.

Audio reasoning is evaluated on BLAB for long-form reasoning (Ahia et al., 2025), MMAR for diverse task coverage (Ma et al., 2025) and MMAU-Pro for diverse skill coverage (Kumar et al., 2025). Our final selection includes 417 problems from BLAB, 888 from MMAR and 3,612 from MMAU-Pro. We analyze Qwen3-Omni-30B as its the only open-weight audio-language model with thinking mode. We exclude commercial models such as Gemini and GPT-4 because they produce summarized traces that are not sufficiently reflective of the underlying reasoning process.

For **image reasoning problems**, we directly use the reasoning traces in Zebra-CoT due to its comprehensive task type coverage and curated reasoning traces. In Zebra-CoT, real-world traces are sourced from online math, physics, coding, and chess datasets. Synthetic traces are created by generating or sourcing images online and writing reasoning templates, then using frontier VLMs (Gemini-2.5 and GPT-4.1) to refine them into more diverse and coherent reasoning traces (Li et al., 2025a). Additionally, this allows us to analyze cognitive elements in synthetic training data beyond raw model reasoning outputs. We sample 1,000 question-reasoning pairs from each task type to obtain 18,000 curated reasoning traces. Across all three modalities, we collect 171,485 model reasoning traces for detailed analysis.

Human Reasoning Data. We collected a small set of human reasoning traces, as qualitative reference points for comparison with LLM-generated reasoning. We recruited 18 human participants to solve a small subset of the GeneralThought dataset while recording their reasoning. These human traces overlap with a subset of the LLM evaluation set and are intended to illustrate how key behaviors manifest in natural human reasoning, rather than to establish a full human benchmark.

Reasoning was recorded using a think-aloud protocol in which they recorded their verbal reasoning (later transcribed with Evernote). Since some tasks require domain-specific facts or state tracking, we allow participants to use *tools*, including web search and note-taking. In such cases, participants were instructed to verbalize the tool usage, e.g., speaking the search keyword aloud.

Each reasoning trace was annotated separately by two different human annotators, giving each reasoning trace a score from a three-level scoring rubric (0=absent, 1=partially present, 2=present) across each behavior in our 28-behavior taxonomy. Scores from different annotators were aggregated via min-pooling, ensuring that estimates of cognitive elements are conservative. These annotation data were later used for iteratively refining the automatic span annotation prompts.

3.1.2 Fine-Grained Cognitive Element Annotation

Each reasoning trace is annotated for the presence of behaviors that embody the cognitive elements from the 28-behavior taxonomy introduced in Section 2 on the span-level. The annotation process identifies specific text segments that demonstrate each cognitive capability, enabling precise localization of behaviors within reasoning processes.

For each cognitive element, we develop annotation guidelines that include: (1) operational definitions grounded in cognitive science literature, (2) concrete behavioral indicators specifying how the capability manifests in text, (3) three-level scoring rubrics, (4) manually curated in-context examples with explanations, and (5) span identification requirements using character indices. The annotation protocol requires marking exact text boundaries using 0-based character positions, ensuring that identified spans can be programmatically extracted and verified.

To ensure psychological precision and annotation consistency, we iteratively refined the prompts human-in-the-loop, using the manual annotation data as the seed (Section 3.1.1) and collecting the same annotators'

²Dilemma type: *resolving situations with contradictory positions and no clear satisfactory solution* (Appendix A.2).

feedback for each round. See Appendix A.1 for exact prompts used for the analysis. Using these refined prompts, full-scale annotation was then performed using **GPT-4.1** with temperature 0.6.

3.1.3 Problem Type Classification & Response Evaluation

Jonassen’s Problem Taxonomy. We extend the problem-solving taxonomy in Jonassen (2000) to classify cognitive tasks in our dataset. Jonassen (2000) defines problem solving as a goal-directed cognitive activity that transforms an *initial state* into a *desired goal state* through systematic reasoning, proposing a typology of 11 problem types: logical problems, algorithms, story problems, rule-using, decision-making, troubleshooting, diagnosis-solution, strategic performance, case analysis, design, and dilemmas. Jonassen & Hung (2015) further characterizes ten out of the eleven problems types along a continuum from well-structured (clear goals, known solution paths, predictable outcomes) to ill-structured (ambiguous goals, multiple solution paths, uncertain outcomes). However, not all tasks in the GeneralThought dataset require goal-directed transformation. We add two categories to capture tasks outside this paradigm: factual recall (retrieving stored knowledge without reasoning) and creative/expressive tasks (generating novel content judged by originality or aesthetic quality rather than convergence to a predetermined solution). This yields a **13-category taxonomy** (Appendix A.2) spanning the full spectrum of cognitive demands in our datasets.

Problem Classification. For each problem, we determine its type through majority voting across three frontier models: 4o-mini, Gemini-2.5-Pro, and Claude-Sonnet-4.5. Each model independently classifies the problem based on the Jonassen (2000) definitions. Three-way disagreements occur in under 3% of cases; we adjudicate these manually using Jonassen’s structural criteria (goal clarity, solution determinacy, domain constraints). This multi-model approach mitigates individual model biases while maintaining scalability.

Response Correctness. AlpacaEval (Dubois et al., 2024) with GPT-4o as the judge is used to assess response correctness. For each problem-response pair, the LLM judge receives the original problem, the model’s response, and a reference response. Ground truth answer is used as the reference for verifiable tasks and Claude-Sonnet-4.5’s response (selected for strong benchmark performance) for non-verifiable tasks.

3.1.4 Cognitive Structure Extraction

Beyond assessing each behavior’s presence within a reasoning trace, our span-level annotation approach enables fine-grained, quantitative analysis of behaviors and their interdependencies. Specifically, we seek to analyze how a reasoner subconsciously structures and sequences specific behaviors throughout their reasoning process. In order to extract this structure, we choose to encode the reasoning trace t as a heterogeneous transition graph G , where each node represents a single behavior from our 28-behavior taxonomy, and each edge can either represent a *hierarchical* (CONTAINS), *sequential* (NEXT), or *parallel* (PAR) behavior relationships. We compute a weight w for each node and edge based on its normalized pointwise mutual information (NPMI) score. We calculate the NPMI using the individual and joint probabilities of **(a)** behavior b occurring in a reasoning trace t and **(b)** trace t successfully resulting in a correct answer.

In order to construct G , we first sort all annotated behavior spans within trace t by their start positions, with ties broken by span length in descending order. For each pair of behaviors (b_a, b_b) with corresponding spans $[s_a, e_a]$ and $[s_b, e_b]$ where $s_a \leq s_b$, we determine their relationship type through a multi-stage classification process:

1. **PAR:** Whether the Manhattan distance between the spans ($|s_b - s_a| + |e_b - e_a|$) falls below a threshold τ_{par} (b_a and b_b occur nearly simultaneously).
2. **CONTAINS:** If $s_b \leq e_a \leq e_b$, we compute the overlap ratio $\rho = \frac{e_a - s_b}{e_b - s_b}$; when ρ exceeds threshold $\tau_{overlap}$ (b_a hierarchically encompasses b_b).
3. **NEXT:** If $e_a < s_b$, $e_a < e_b$ and the overlap is below $\tau_{overlap}$, or if $e_a > e_b$ but their overlap ratio $\frac{e_b - s_b}{e_a - s_a} < \tau_{overlap}$ (b_a is followed by b_b).

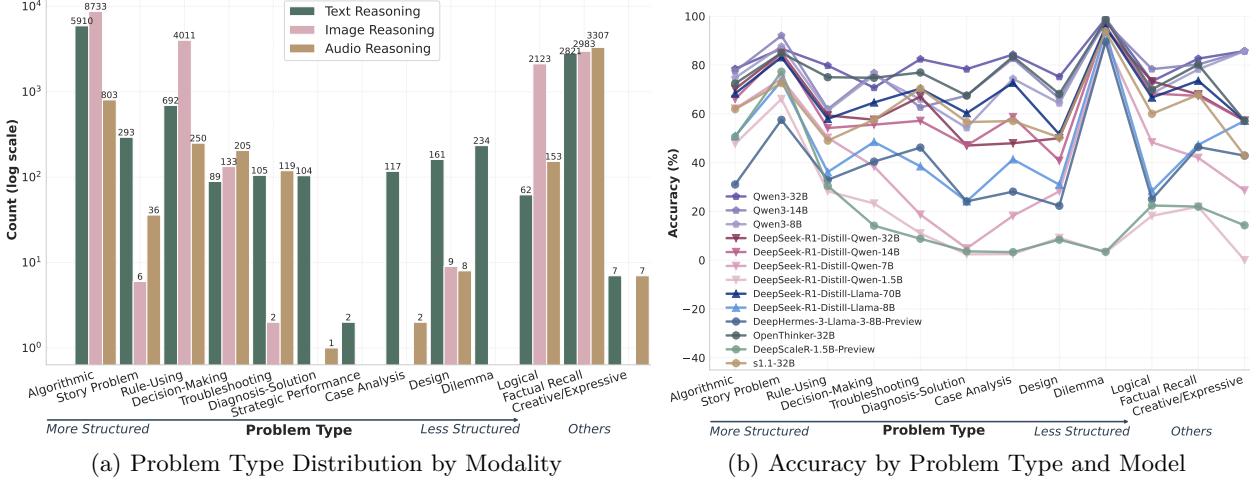


Figure 3: Dataset composition and model performance across problem types. (a) Problem distribution across modalities, organized by Jonassen’s problem structuredness continuum, shows coverage decreases for less structured problems. (b) Accuracy decreases as problems become less structured, with models showing consistent performance on algorithmic tasks (63.8%) but high variance on dilemmas (3.3-99.1%).

We set $\tau_{par} = 20$ (characters) and $\tau_{overlap} = 0.8$. To ensure graph sparsity and capture only direct sequential dependencies, we apply a refinement step: for each behavior b_a , we retain only the first NEXT edge to an immediately subsequent non-overlapping behavior, filtering out transitive connections to more distant behaviors. This process yields a directed graph that preserves both the hierarchical decomposition and temporal ordering of reasoning behaviors within each trace.

3.2 Experimental Setup

3.2.1 Dataset Composition

Problem Type. The distribution of problem types (Figure 3a) varies substantially across modalities but decrease as the problem becomes less structured. Algorithmic problems dominate with 6,300 text, 8,400 image, and 800 audio samples. Among structured types, Rule-Using shows strong concentration in images (4,447 samples) versus text (521) and audio (248). Notably, Strategic Performance has zero instances across all modalities, while the LLM classifiers frequently considered this category with explanations including “*I considered Strategic Performance but rejected it because this is an offline planning problem under perfect information, not real-time execution.*” Strategic Performance necessitates interactive benchmarking, which is not feasible under our experimental constraint to generate reasoning traces to the same input across models, but the lack of this problem type in the single-turn, static benchmarks echoes the call for interactive evaluation of models (Hofmann et al., 2025; Li et al., 2024). Three problem types fall outside the structuredness ranking: Factual Recall is the most prevalent, showing balanced representation across modalities (2,841 text, 2,663 image, 3,307 audio) and representing the second-largest category overall. Logical problems, involving abstract reasoning, are heavily image-focused (2,215 samples) but rare in text (57) and audio (152). Overall, text reasoning shows the broadest distribution across problem types, while image and audio concentrate on Algorithmic, Rule-Using, Logical, and Factual Recall.

Success Rate. Success rates vary substantially by problem structure (Figure 3b). Well-structured problems achieve higher correctness overall (algorithms: 63.8%, story problems: 79.5%), while ill-structured problems average lower success (design: 46.6%). Dilemma tasks show polarization (3.3-99.1%) depending on the model size. Detailed model-specific performance breakdowns appear in Appendix A.3.

3.2.2 Analysis Dimensions

To thoroughly examine the manifestation of behaviors within reasoning traces, we organize our analysis around three interconnected research questions:

1. **Which behaviors are most prevalent in reasoning traces, and how does their frequency relate to reasoning success?** We investigate the distribution of reasoning behaviors across different models and problem types, examining whether prevalence patterns vary along Jonassen’s well-structured to ill-structured problem continuum. Critically, we distinguish between behaviors that are frequently exhibited and those that are most strongly correlated with correct outcomes, as these sets may not align. This analysis reveals whether models consistently employ the behaviors most conducive to success or instead are biased towards alternative strategies.
2. **What structural dependencies exist between reasoning behaviors?** Beyond examining individual behavior presence, we analyze their temporal and hierarchical interdependencies within reasoning traces. The mere presence of success-correlated behaviors does not guarantee effective reasoning; rather, the *ordering and composition* of behaviors critically influences problem-solving efficacy. For instance, when resolving a dilemma, failure to employ conceptual processing to decompose the problem into constituent considerations early in the reasoning process may fundamentally compromise the efficiency and quality of the eventual judgment, even if these behaviors appear later. To identify optimal behavior sequences, we leverage our transition graph representation (Section 3.1.4) to extract consensus subgraphs—common structural patterns that maximize collective NPMI scores across successful traces, thereby revealing the behavioral scaffolding most strongly associated with correct reasoning.
3. **How do behavioral structures differ between LLMs and human reasoners?** We conduct a comparative study examining both the distributional characteristics of behaviors exhibited by LLMs versus humans on a shared problem set, and perform fine-grained, span-level qualitative analysis to understand how they differ in their reasoning processes and behavior utilization strategies.

3.3 Results & Analyses

3.3.1 Distribution of Cognitive Elements

The first phase of our analysis investigates the misalignment between cognitive elements that models frequently deploy (through behavioral manifestation) versus those that correlate most strongly with success. Figure 4 presents this contrast: the left heatmap shows behavioral prevalence across all traces for each problem type, while the right displays the positive pointwise mutual information (PPMI) between behavior occurrence and trace success.

Models deploy behaviors inversely to what success requires. The contrasting patterns between the two heatmaps reveal a fundamental misalignment in how models adapt their reasoning strategies. On well-structured problems (e.g., algorithmic, rule-using), models deploy a broad repertoire of behaviors at high frequencies, where the average presence across all behaviors is 0.3652 ± 0.2598 for algorithmic, story, and rule-using problems. As problems become ill-structured and non-verifiable, however, models narrow their behavioral repertoire: ill-structured problems like case analysis, design, and dilemma have a 0.2787 ± 0.2672 average presence across all behaviors, with usage concentrating heavily on sequential organization, logical coherence, and forward chaining.

This narrowing strategy is the opposite of what is reflected in successful traces. The right heatmap reveals that *overall success on ill-structured problems demands greater behavioral diversity*. As we move from well-structured to ill-structured problem types, the PPMI scores become more uniformly elevated across a wider range of behaviors, particularly diverse representations (hierarchical, network, spatial, temporal) and varied operations (backward chaining, representational restructuring, pattern recognition). Specifically, algorithmic, story, and rule-using problems have an average PPMI score of 0.052 ± 0.067 across all behaviors, while the last four ill-structured problems (from diagnosis-solution to dilemma) have an average score of 0.224 ± 0.182 .

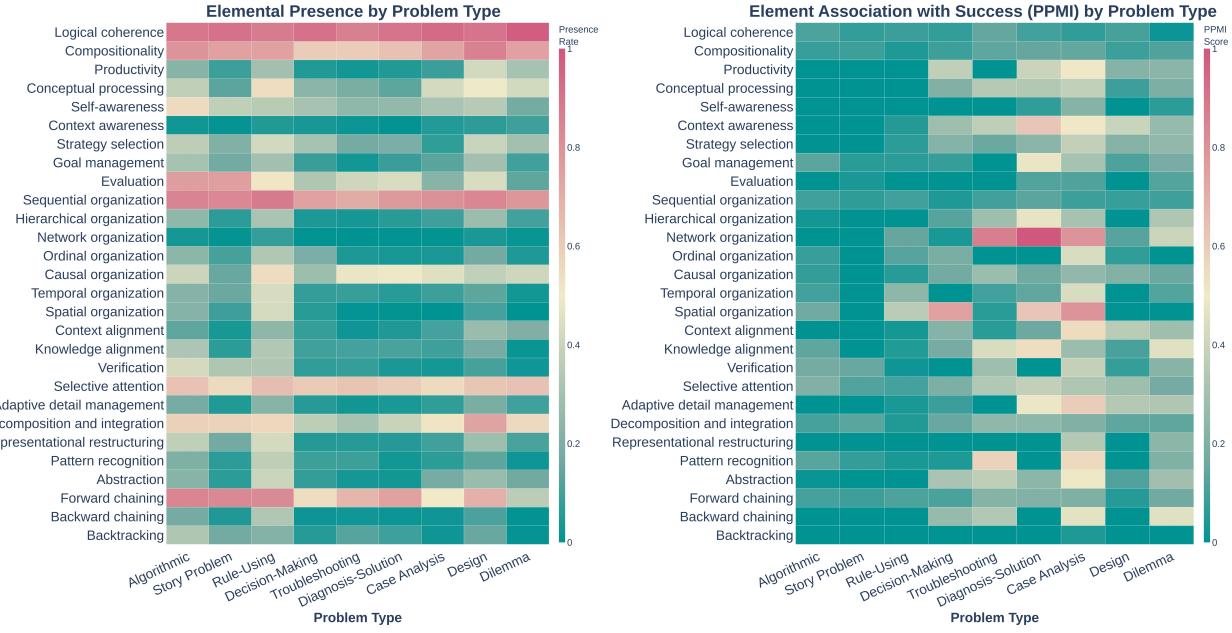


Figure 4: (**Left**) Presence rate of each reasoning behavior for each problem type (ranging from well-structured to ill-structured). (**Right**) Positive Pointwise Mutual Information (PPMI) between the problem type and reasoning behavior (correlation between the behavior occurrence and reasoning trace success).

Thus, while well-structured problems, which are more tolerant of uniform approaches, receive broad behavioral engagement, ill-structured problems, which critically require a diverse repertoire of reasoning behaviors, only receive a subset dominated by `sequential processing` and `forward chaining`. This inverse relationship between behavioral deployment and success reflects that models have learned to apply their most diverse reasoning behaviors where they are the least necessary, *while defaulting to limited, inflexible strategies precisely where adaptability matters most*.

Models frequently attempt core reasoning behaviors but struggle to execute them effectively. Reasoning invariants—particularly `logical coherence` and `compositionality`—appear ubiquitously across problem types, yet show surprisingly weak correlation with success. For example, logical coherence’s average NPMI across all problem types is 0.0984, whereas knowledge alignment features an average score of 0.301. Manual inspection of traces by human annotators reveals a systematic pattern. While models frequently attempt to identify logical inconsistencies and contradictions, they consistently fail to recognize or effectively respond to them, unlike human reasoners. This execution gap explains the discrepancy between high prevalence and low predictive value of these foundational behaviors.

Models show limited meta-cognitive success, especially on problems lacking clear ground truth. The `evaluation` element similarly demonstrates high prevalence but negligible success correlation, with its frequency declining sharply for ill-structured problems (case analysis, dilemma). This pattern suggests models struggle particularly with self-assessment on non-verifiable problems where ground truth is ambiguous or absent.

Default sequential organization hampers performance when problems require alternative representations. Models exhibit strong preference for `sequential` and `causal organization` regardless of problem type. However, the success patterns indicate that as the problem type becomes less structured (algorithmic → dilemma), diverse organizational strategies—including hierarchical, network, ordinal, temporal, and spatial representations—become increasingly critical. This reflects a fundamental challenge of ill-defined problems: when problem descriptions lack inherent structure, successful reasoners must actively construct appropriate organizational frameworks rather than defaulting to sequential processing.

Operational rigidity limits models’ ability to adapt reasoning strategies to problem demands. Similar rigidity emerges in reasoning operations. Models consistently favor `selective attention`,

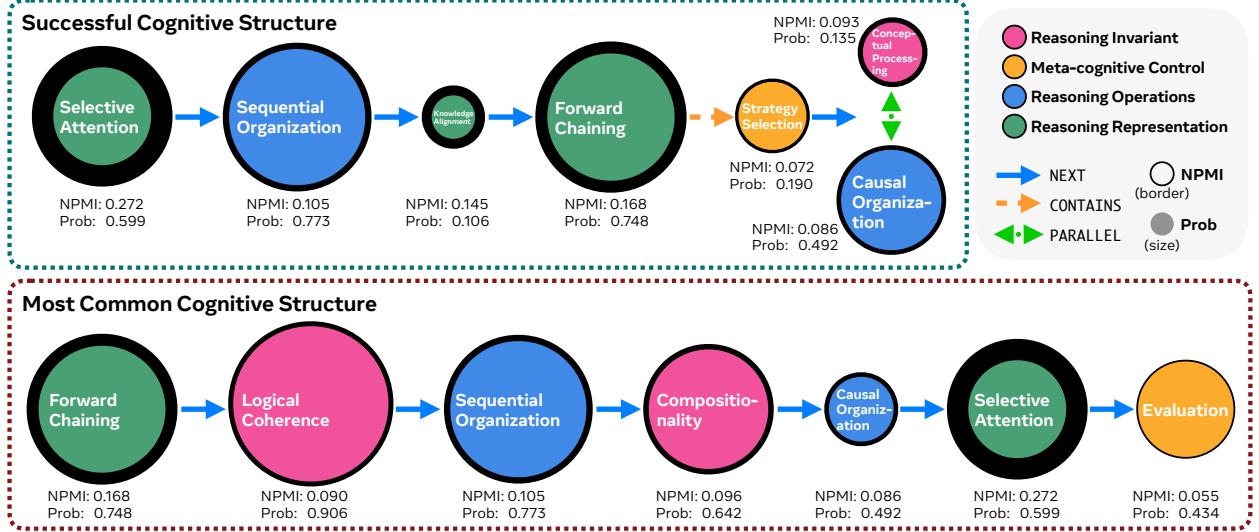


Figure 5: The *most successful* vs. *most common* 7-node behavioral pattern across all *Diagnosis-Solution* reasoning trace with the correlation to success (NPMI) and the probability of each node.

decomposition and integration, and **forward chaining**, the latter being a natural consequence of their sequential organizational bias. Successful traces, however, demonstrate substantially greater operational diversity, adapting their reasoning strategies to problem characteristics rather than applying uniform approaches.

These findings reveal a fundamental misalignment: models seem to deploy behavioral strategies based on learned priors rather than adapting to problem-specific demands, resulting in systematic gaps between frequently used behaviors and those that actually drive success.

Model-Specific Behavior Distribution To complement our problem-type analysis, we examine how behavioral deployment varies across different model architectures and modalities. Figure 6 displays the presence rate of each reasoning behavior across eleven models, including text-only systems (the majority), audio-capable models (Qwen3-Omni), and vision-language models (Zebra-CoT).

Several behaviors exhibit consistently high presence rates across nearly all models, suggesting core reasoning primitives that emerge regardless of architecture or training methodology. **Backtracking** appears frequently across all models (presence rates exceeding 0.8 across all models), indicating that iterative refinement and error correction are prominent components of model reasoning processes.

With respect to the different modalities, we observe that the audio and image models have overall less presence of reasoning behaviors, exhibiting certain behaviors frequently (e.g., **evaluation**, **productivity**) and others minimally, relative to the textual models. This indicates that there may be lower diversity of reasoning behaviors present in multimodal models.

3.3.2 Behavioral Structures

Our analysis in Section 3.3.1 demonstrated that successful reasoning traces exhibit diverse behaviors, adapting their selection to problem-specific demands. However, behavioral presence alone provides an incomplete picture. Thus, we ask: does the *temporal and hierarchical organization* of these behaviors also differ between successful and unsuccessful traces? To address this question, we extract optimal behavioral structures for each problem type by analyzing the transition graphs constructed from all model traces T .

Behavioral consensus subgraph extraction. We leverage the graph construction methodology detailed in Section 3.1.4 to encode and aggregate reasoning traces for each problem type, yielding a complete transition graph G where each node represents a behavior and each edge captures their relationship (hierarchical,

sequential, or parallel). Each node and edge is weighted by its NPMI score, quantifying the strength of association between that behavioral element and trace success.

Given graph G , we extract a consensus subgraph $G^* \subseteq G$ that represents the prototypical successful reasoning structure through a greedy forward search procedure. First, we identify the starting behavior by selecting the node with highest NPMI among behaviors that initiate successful traces (i.e., behaviors with the earliest temporal spans). We denote this initial node as b_{start} . From the current behavior b_{curr} , we iteratively expand the consensus structure by selecting the outgoing edge $(b_{\text{curr}}, b_{\text{next}}, e_{\text{type}})$ with the maximum NPMI score, where $e_{\text{type}} \in \{\text{CONTAINS}, \text{NEXT}, \text{PAR}\}$. To ensure acyclicity and prevent redundant paths, we maintain a visited set and exclude any target behaviors already present in G^* . This process continues until either: (1) we reach a predefined maximum number of nodes $|V_{\text{max}}|$, (2) no valid outgoing edges exist from b_{curr} , or (3) when operating in dynamic mode, all remaining edges have non-positive NPMI scores, indicating they do not positively correlate with success.

The resulting consensus subgraph G^* represents the behavioral scaffolding most strongly associated with successful reasoning for the given problem type. Critically, this structure captures not merely which behaviors appear, but *how they are sequenced and composed*—revealing the temporal dependencies and hierarchical decompositions that characterize effective problem-solving strategies. By comparing G^* across problem types, we can identify whether optimal reasoning structures remain consistent or adapt fundamentally to different levels of problem structure.

Analysis. Figure 5 illustrates the consensus subgraph constructed for Algorithmic problems. We observe that successful traces begin with selective attention to identify relevant problem features, followed by sequential organization and forward chaining to execute solution steps. This structure contrasts sharply with the most common pattern, which prioritizes logical coherence and compositionality early but achieves lower success correlation. The divergence suggests that immediate focus on abstract reasoning principles without first attending to problem-specific constraints leads to less effective problem-solving on algorithmic tasks.

3.4 Comparison with Humans

Distribution of Elements. To complement our analysis of cognitive element presence and structure across models and problem types, we compare 30 manually annotated reasoning traces from both humans and LLMs. Figure 7 presents elements annotated as *strongly* present, excluding partial occurrences. Human traces exhibit significantly higher rates of abstract cognitive elements, such as self-awareness (49% vs. 19%), abstraction (54% vs. 36%). Notably, reasoning representations were frequently marked as partially present across both groups, particularly for humans, likely reflecting the more implicit nature of their internal knowledge representations and reasoning processes. Conversely, LLMs demonstrate greater reliance on backward chaining (navigating a problem by reasoning from goals to prerequisites; less common in human problem-solving), and higher productivity scores (potentially attributable to LLMs’ tendency to externalize intermediate reasoning steps that humans often leave implicit).

Qualitative comparative analysis. Table 3 illustrates the difference between human and LLM reasoning traces. As shown, Humans are quicker to invoke conceptual processing and abstraction in logical problem solving, leading to significantly shorter reasoning traces. LLM’s on the other hand, often resort to surface level

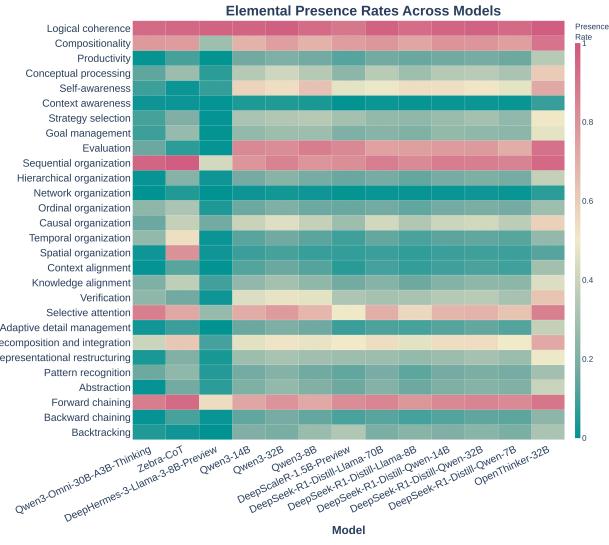


Figure 6: Presence rate of each reasoning behavior for each model (how often is the behavior occurring across all reasoning traces for a model), including all modalities (Qwen3-Omni: audio; Zebra-CoT: image).

reasoning principles without first attending to problem-specific constraints leads to less effective problem-solving on algorithmic tasks.

Table 3: Illustrative examples of differences between human and LLM reasoning traces. This table shows partial (for brevity) reasoning traces of both humans and DeepSeek-R1 on the same problems for both Logical and Design problems. Spans containing behavioral manifestation of cognitive elements, or lack thereof are annotated, with the cognitive element added using <> tags.

Problem	Annotator	Reasoning Trace Excerpt
Logical	Human (166 tokens)	... So here's how we do it. It is eight square right and we painted the board with black and white just like, you know, chessboard. I'm not sure but what you will find out is that the two corner you throw away is both black or... Anyway, let's first paint the upper left corner black and any cell near it white. And then the cell near white cell is black and so on and what you will find is that the two... the two corners are always both black <abstraction> . But every domino covers exactly white and black. So you cannot do that because you have more white cells than the black cells <conceptual processing> . Okay. So this is it, I think <no verbalized verification> .
	DeepSeek-R1 (7,632 tokens)	... In an 8x8 checkerboard, each diagonal has squares of alternating colors. ... Wait, let's count. (1,1) is black, (2,2) is white, (3,3) black, (4,4) white, (5,5) black, (6,6) white, (7,7) black, (8,8) white <no abstraction> . So the diagonal from (1,1) to (8,8) would have squares that alternate colors. [...] Wait, maybe the diagonally opposite corners are the same color <slight abstraction> . Let me check again. [...] Wait, maybe there's another reason <verification> . Maybe the parity or the arrangement. [...] For example, if the board is split into regions that can't be covered by dominoes. But how?... Therefore, even though the black and white counts are balanced, the parity based on i+j is not, leading to impossibility.
Design	Human (2,071 tokens)	For this task is to decide on a comprehensive policy reform for healthcare that integrates three goals. <decomposition> ... And the approach I'm going to take here is <strategy selection> ... I will pick the system that has the best rank overall over those three trade-offs. <ordinal organization> ... So searching with Google, ... the second result I get is from Columbia University, their school of public health. So that seems like a reliable source. <context awareness> And it says, I go to this page, it says types of healthcare systems. <selective attention> It says there are four different types of healthcare models. All right, so that sounds pretty promising. <evaluation> ... I'm gonna say one aspect of quality of care is access. ... I think the one where the government pays but you get to pick your own doctor, that was going to have the best access. So I'm going to say the single payer model is going to be number one. ... And then in the Bismarck model, there's still public and private actors. So I'm going to give that number two, and then Beveridge is going to get number three. <knowledge alignment> ... So what I'm going to do is I'm going to calculate the composite score by summing the ranks. <abstraction> ... And so this surprised me by my ranking out of pocket is the best system. You know, ... where you pay your doctor a fee directly ... without an insurance company. So this surprises me a bit. <self-awareness>
	DeepSeek-R1 (747 tokens)	Okay, so I need to come up with a comprehensive policy reform for the U.S. healthcare payment system. The goal is to integrate payment systems to reduce waste and fraud, improve quality, and make healthcare more patient-centered. Let me start by breaking down the problem. <decomposition> First, the current system is fragmented. There are multiple payers—Medicare, Medicaid, private insurers, etc.—each with their own payment structures. This fragmentation probably leads to administrative inefficiencies. Providers have to deal with different billing systems, which is time-consuming and costly. Also, patients might get confused with varying coverage and out-of-pocket costs. <causal organization> ... So, putting this all together, the policy should have components like standardized payment systems, value-based reimbursement, centralized fraud detection, patient empowerment through transparency, and integration of social services. Also, regulatory support and phased implementation to manage the transition.

reiteration and enumeration. While LLMs try to reason abstractly and verify their reasoning, our Human annotators have repeatedly observed that they fail at learning from previous verifications in subsequent problem solving attempts, and they often repeat verification and backtracking on claims and paths that have already been explored. On more open ended questions, or questions that require complex factual recall, humans tend to invoke higher order behaviors such as self-awareness and strategy selection, leading to longer traces than LLMs, which seem to rely more on factual recall of relevant information.

4 Eliciting Cognitive Reasoning Structures

Section 3.3.2 established a methodology for extracting consensus subgraphs that encode the optimal hierarchical and temporal sequencing of reasoning behaviors for each problem type. We now investigate whether these empirically-derived cognitive structures can provide effective *test-time reasoning guidance*, steering models toward statistically successful reasoning patterns and thereby improving task performance.

Methodology. To operationalize cognitive structure guidance, we adopt a straightforward approach: automatically converting each problem type’s consensus subgraph into an actionable prompt that contextualizes the reasoning structure and explicitly scaffolds the model’s problem-solving process. We first generate a lin-

Table 4: Model performance improvement after steering using cognitive structure guidance (nodes in graph = 7). Values show percentage change: $\frac{\text{After} - \text{Before}}{\text{Before}} \times 100\%$. Approximately 50 problems per model and problem type, sampled equally from previously incorrect and correct answers. Color intensity indicates magnitude: darker green for larger improvements, darker red for larger degradations. Design problems showed catastrophic failure (0% accuracy) across all models after steering and are excluded.

Model	Avg.	Log	Algo	Story	Rule	Decision	Troub	Diag	Case	Dilem	Fact
Hermes-3-Llama-3-8B	-17.5	+0.0	-72.0	-68.0	-72.0	-4.0	-8.0	+41.2	+12.0	+56.0	+44.0
DeepScaleR-1.5B	-52.6	-71.4	-56.0	-87.0	-23.9	-50.0	-57.1	+0.0	+14.3	-33.3	-48.0
R1-Distill-Llama-70B	+21.9	+26.8	+12.4	+18.8	+36.0	+28.0	+36.0	+48.0	+48.0	+54.1	+28.0
R1-Distill-Llama-8B	+4.0	+16.8	+8.4	+36.0	+0.0	-12.0	-4.0	+16.7	+36.0	+48.0	+4.0
R1-Distill-Qwen-14B	+14.6	+31.1	+50.0	+20.1	+28.0	+28.0	+12.0	+20.0	+28.0	+40.0	+0.0
R1-Distill-Qwen-32B	+22.3	+15.6	+36.0	+12.5	+40.0	+20.0	+32.0	+36.0	+56.0	+60.0	+40.0
R1-Distill-Qwen-7B	-6.3	+8.4	-12.4	+28.0	+24.0	+0.0	-31.2	+66.7	+4.0	+0.0	-4.0
OpenThinker-32B	-1.3	-9.4	-12.0	-30.0	+28.0	+12.0	+4.0	+6.9	+29.6	+37.5	+12.0
Qwen3-14B	+22.5	+16.2	+32.0	+7.5	+44.0	+20.0	+44.0	+50.0	+52.0	+60.0	+32.0
Qwen3-32B	+14.8	+11.8	+24.0	+20.6	+20.0	+24.0	+15.2	+24.2	+41.9	+19.5	+48.0
Qwen3-8B	+13.8	+3.1	+36.0	+11.1	+24.1	+16.0	+28.0	+48.0	+36.0	+7.9	+48.0

earized representation of each behavioral graph, then apply an automated prompt construction procedure to produce test-time guidance instructions (examples provided in our codebase). This fully automated pipeline, requiring no expert prompt engineering, allows us to assess whether models can leverage structural guidance without hand-crafted templates.

We evaluate this approach on a stratified sample of approximately 50 textual problems per problem type, deliberately balancing between questions the model previously answered correctly and incorrectly. This sampling strategy serves two critical purposes: (1) verifying that guidance does not degrade performance on problems the model already solves successfully, and (2) measuring improvement on previously failed instances. We quantify effectiveness by computing the percentage change in accuracy after applying cognitive structure guidance relative to the model’s baseline performance.

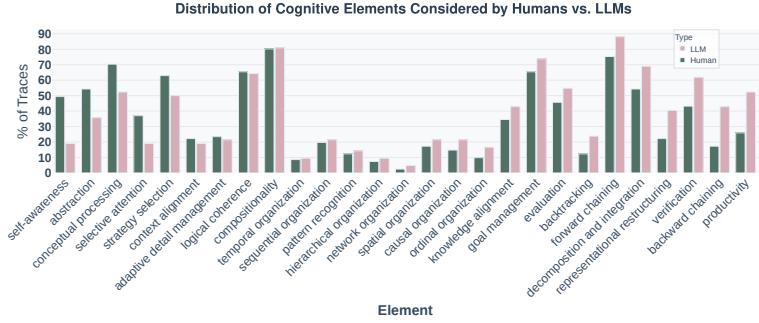


Figure 7: Presence rate of each reasoning element for humans and randomly sampled LLMs. We filtered the presence judgments for *strongly present* elements. X-axis is sorted based on the presence gap between humans and LLMs.

Results and Analysis. Table 4 reveals substantial heterogeneity in how models respond to cognitive structure guidance. Modern, capable reasoning models (particularly the Qwen3 family and larger R1-Distill variants) demonstrate significant improvements, with gains reaching up to 60% on ill-structured problem types (e.g., +60.0% for Qwen3-14B on dilemmas, +60.0% for R1-Distill-Qwen-32B on dilemmas). Notably, these improvements concentrate most strongly on complex, open-ended problems (dilemmas, case analysis, diagnostic reasoning) where explicit structural scaffolding appears most beneficial.

However, this effectiveness is highly dependent on model capacity and architectural sophistication. Smaller or less capable models, particularly Hermes-3-Llama-3-8B and DeepScaleR-1.5B, show pronounced performance *degradation* across most problem types, with losses exceeding 50% in several categories (e.g., -72.0% for both models on algorithmic problems). This suggests a capability threshold: models must possess sufficient reasoning flexibility and instruction-following ability to productively adapt their processes to detailed structural guidance. Below this threshold, explicit scaffolding appears to constrain rather than enhance rea-

soning, potentially by overwhelming limited cognitive resources or conflicting with ingrained problem-solving heuristics.

The pattern of improvements also illuminates problem-type specificity. Well-structured problems (algorithmic, rule-using) show more modest or even negative effects across models, while ill-structured problems exhibit the strongest positive responses in capable systems. This asymmetry likely reflects two factors: (1) well-structured problems may already elicit relatively optimal reasoning patterns from models, leaving less room for improvement, and (2) ill-structured problems benefit more from explicit organizational scaffolding that helps models navigate ambiguous problem spaces.

Collectively, these findings provide strong preliminary evidence that **optimal cognitive reasoning structures aligned to problem characteristics can substantially enhance model performance**—but only when models possess the architectural sophistication to leverage such guidance effectively. This suggests a promising direction for adaptive reasoning systems that dynamically apply problem-type-specific cognitive scaffolding to capable foundation models.

5 Cognitive Element Considerations in LLM Research Design

To contextualize the behavioral presence observed above in Section 3.3.1, we examine how contemporary LLM research conceptualizes "reasoning" as a design choice. With growing interest in the development and analysis of models with "diverse and strong reasoning" abilities, we aimed to understand what cognitive elements are currently supported and which remain underexplored. For this analysis, we scraped arXiv³ papers using keyword-based queries applied to titles and abstracts. We searched using general reasoning-related keywords (e.g., "large language model reasoning," "LLM cognitive behaviors," "LLM thinking") and for each capability in our taxonomy we added an additional query of the form "large language model <capability>", retrieving the top ten papers for each general query (in order to maximize precision) and 100 papers for each behavior-specific query.

We then annotated this set of papers using both two human annotators and GPT-4.1 (all three achieve moderate agreement $ICC_{3k} = 0.593$; we iteratively improve annotation prompts based on human feedback). For each paper–capability pair, we record whether a behavior is *explicitly targeted* (via evaluation objectives or architecture), *implicitly encouraged* (dataset structure, demonstration style), or *not incorporated*.

Results from this analysis displayed in Figure 2 reveal substantial imbalance in the types of reasoning behaviors emphasized across LLM reasoning research. The dominant cognitive elements (context awareness: 70% of papers, decomposition and integration: 60%, knowledge structure alignment: 56%) align with linear procedural step-by-step reasoning that is straightforward to evaluate. Sequential organization (54%), pattern recognition (51%), and abstraction (46%) similarly emphasize forward-moving compositional patterns. In contrast, behaviors enabling flexible problem-solving receive far less attention: self-awareness appears in only 16% of papers, spatial organization in 10%, and temporal organization in 22%.

Comparing paper-level design intentions to observed model behaviors reveals three key discrepancies. First, **design-behavior gaps**: Compositionality appears in 38% of papers yet shows inconsistent presence in reasoning traces. Context alignment and knowledge structure alignment are frequently targeted (47% and 56%) but models struggle to maintain these consistently across problem types. Second, **emergent but under-theorized behaviors**: Profundity appears in only 16% of papers but manifests consistently in model outputs, suggesting models develop meta-cognitive patterns not explicitly designed for. Third, **systematically neglected capabilities**: Self-awareness (16%), temporal organization (22%), ordinal organization (27%), and spatial organization (10%) represent sophisticated cognitive structures essential for non-linear thinking, mental simulation, and metacognitive monitoring. These capabilities receive minimal research attention and fail to emerge spontaneously in model behavior.

This synthesis reveals a critical bottleneck. Current LLM reasoning research operates within a narrow conceptual vocabulary privileging linear, compositional behaviors amenable to straightforward evaluation.

³<https://arxiv.org/>

Expanding theoretical foundations beyond step-by-step decomposition toward richer cognitive taxonomies may be essential for developing more sophisticated reasoning capabilities.

6 Opportunities and Challenges

Our proposed cognitive foundations taxonomy enables systematic characterization of reasoning processes, allowing us to answer questions like *which cognitive elements appear in models?*, *how behaviors unfold during reasoning?* and *which behavioral patterns correlate with success?* Overall, our analyses expose fundamental gaps: we cannot know which training produces which cognitive capabilities *a priori*, cannot ensure behaviors transfer beyond training distributions, and cannot validate whether observed patterns reflect genuine cognitive mechanisms or spurious reasoning shortcuts. The cognitively inspired test-time reasoning guidance demonstrates that this understanding is actionable, in which successful behavioral patterns can be elicited to improve performance on ill-structured problems (Table 4). However, even in this case, it remains unclear whether our guidance enables genuine deployment of latent capabilities or simply helps models retrieve cached reasoning patterns from training data.

Cognitive science research provides principled frameworks for diagnosing gaps in current systems and designing interventions to address them. We posit that decades of research on problem-solving, mental representation, and meta-cognition offer concrete guidance for technical development of LLMs. The following challenges illustrate how cognitive theories illuminate current limitations and suggest specific research directions.

Predicting cognitive capabilities from training procedure. Current training paradigms lack predictive theories connecting procedures to emergent reasoning capabilities. Meta-cognitive monitoring correlates strongly with success on ill-structured problems (Figure 4), yet appears in only 8% of reasoning traces and 8% of papers (Figure 2). Our comparison across 16 models shows dramatic variation in response to cognitive scaffolding (Table 4), but we cannot predict these differences from architecture or training details. Post-hoc analyses reveal that RL induces verification (Gandhi et al., 2025; Snell et al., 2024) but not representational restructuring or meta-cognitive monitoring, while process supervision (Lightman et al., 2023; Uesato et al., 2022) and chain-of-thought prompting (Wei et al., 2022b; Kojima et al., 2023; Wang et al., 2022; Yao et al., 2023) elicit latent but not spontaneous behaviors. We cannot explain why RL produces verification but not meta-cognitive monitoring, or whether architectural prerequisites exist for specific behaviors versus emerging from scale (Le et al., 2025; Das et al., 2025).

Cognitive science provides explanatory frameworks. Research on skill acquisition shows different capabilities emerge under different learning conditions (Flavell, 1979; Nelson, 1990). Procedural skills like verification emerge from repeated task performance, while meta-cognitive monitoring requires explicit reflection on reasoning processes (Nelson, 1990). Domain-specific strategies require rich within-domain examples, while transferable schemas require diverse, structurally varied training (Gentner, 1983).

These principles enable predictive theories of capability emergence. If procedural behaviors emerge from repetition while meta-cognitive behaviors require reflection, we can predict which training paradigms will produce which capabilities before running experiments. If transferable reasoning requires structural variation across surface features, we can predict that training on homogeneous problem distributions will fail to produce robust generalization. Our taxonomy provides measurement infrastructure for testing these predictions; cognitive theories provide the explanatory framework linking training characteristics to behavioral outcomes. This enables shifting from post-hoc observation to theory-driven experimentation.

The generalization challenge. Reasoning behaviors fail to transfer beyond training distributions. Models achieve 80% on story problems but 46% on design problems (Figure 3b). They rely on shallow forward chaining that fails when problems demand hierarchical planning and representational restructuring. Our test-time reasoning guidance improves performance through automatically constructed cognitive structure templates contextualized to each problem type. However, it still requires prior knowledge of a diverse distribution of patterns and their success outcomes. This brittleness reflects a fundamental finding from cognitive science: transfer depends on abstract schema formation. Human reasoning transfers through schemas that

capture structural commonalities across surface-different problems (Gentner, 1983; Gick & Holyoak, 1980). This requires representing problems at multiple levels of abstraction, recognizing when novel situations instantiate known structures, and flexibly adapting strategies (Holyoak & Thagard, 1997). Children develop transferable mathematical understanding through varied examples that highlight structural principles (Rittle-Johnson et al., 2001). Adults spontaneously recognize deep structural similarity despite surface differences (Novick & Hmelo, 1994). Crucially, transfer fails when learners encode only surface features without extracting underlying structure (Gick & Holyoak, 1980).

LLMs exhibit precisely the failure modes predicted by cognitive theory by succeeding in-distribution but fail on superficial variants (McCoy et al., 2019; Berglund et al., 2023; Shi et al., 2022; Shao et al., 2025; Li et al., 2025b). If models only deploy desired behaviors under explicit prompting, they may be applying cached patterns rather than reasoning about which strategy the problem demands which is exactly the behavior embodied by surface-level learning without schema abstraction in humans. Applying principles of schema-based transfer to training offers a potential technical solution. Cognitive research demonstrates that transfer improves when: (1) training highlights structural similarities across diverse surface forms, (2) learners explicitly compare and contrast examples to extract common structure, (3) training includes prompts to reflect on *why* a strategy worked (Gentner, 1983; Gick & Holyoak, 1980). For models, this suggests training procedures that explicitly encourage structural comparison across problem types, reward strategy selection based on problem structure rather than surface features, and potentially use contrastive learning to distinguish structural from surface similarity. Our framework enables measuring whether these interventions produce genuine transferable understanding (flexible behavioral deployment across problem types) or spurious reasoning shortcuts (rigid strategies that succeed only in-distribution).

From observable behavior to underlying processes. Our framework identifies observable behavioral patterns, but the same surface behavior can arise from fundamentally different underlying processes (Chomsky, 2014). Our human-LLM comparison illustrates this: both reach correct answers, yet humans employ hierarchical nesting and meta-cognitive loops while models use shallow sequential chaining (Figure ??). Recent work shows models produce correct reasoning chains while internally representing different processes (Turpin et al., 2023) and that chain-of-thought may be post-hoc rationalization (Lanham et al., 2023). These alternatives produce similar outputs but fundamentally different robustness and generalization.

Cognitive science research on reasoning invariants provides principled validation criteria. Genuine cognitive capabilities exhibit systematic transfer to structurally similar but surface-different contexts, robustness to perturbation in irrelevant dimensions, compositional deployment across tasks, and internal consistency when combined with other capabilities (Fodor & Pylyshyn, 1988; Gentner, 1983). These are functional properties requiring testing under manipulation, not mere behavioral observation.

Developing validation frameworks that test these signatures requires moving beyond "does behavior X appear?" to "does X transfer systematically, remain robust to perturbation, and compose flexibly?" This demands systematic probing studies measuring transfer and robustness (Belinkov, 2022), causal interventions manipulating specific capabilities while measuring downstream effects, and mechanistic interpretability connecting internal representations to behavioral function (Elhage et al., 2021; Nanda et al., 2023). Our taxonomy identifies which behaviors to validate; cognitive theories specify what properties distinguish genuine capabilities from spurious shortcuts.

Expanding behavioral coverage and diversity. Our analysis of 1,598 LLM reasoning papers (Figure 2) reveals concentration on easily quantifiable behaviors: sequential organization and decomposition dominate (55%, 60%) while meta-cognitive controls receive minimal attention (self-awareness: 16%, evaluation: 8%). Yet our empirical findings show diverse behavioral repertoires correlate with success on ill-structured problems where rigid strategies fail (Figure 4), consistent with cognitive science findings that meta-cognitive monitoring enables error detection (Nelson, 1990), representational flexibility predicts success on complex problems (Ohlsson, 1992), and sophisticated operations distinguish experts from novices (Chi et al., 1981).

Current RL training faces a fundamental reward specification problem: outcome-based rewards (Lambert et al., 2024) provide sparse, terminal signals that fail to incentivize intermediate reasoning behaviors predicting success (Li et al., 2025b). Process reward models (Lightman et al., 2023; Uesato et al., 2022) reward

intermediate correctness but still optimize for accuracy rather than behavioral diversity enabling transfer. Moreover, RL exploration typically relies on distributions learned from pretraining and midtraining that may fail to discover diverse reasoning strategies (Shao et al., 2025; Olmo Team et al., 2025). Our cognitive taxonomy offers a structured approach to bootstrap exploration by explicitly targeting underexplored behavioral patterns. For instance, seeding rollouts with prompts encouraging backward chaining when models predominantly use forward chaining, or constraining generation to employ spatial organization when sequential organization dominates.

Three technical opportunities emerge. First, *reward shaping for meta-cognition*: augment reward models to evaluate error detection, strategy adaptation, and explicit reasoning about problem structure through multiple objectives or explicit architectural modifications (Li et al., 2025c). Second, *curriculum design for representational diversity*: developmental psychology shows children acquire flexible reasoning through structured variation in problem presentation (Rittle-Johnson et al., 2001), which can be operationalized in LLMs through multi-task RL with auxiliary tasks rewarding specific organizational structures, or curricula that progressively require different representations of the same underlying structure. Third, *environment design penalizing brittleness*: procedurally create training distributions where rigid strategies fail—problem variants requiring backward chaining after forward-chaining training, or adversarial problems exposing shallow heuristics (Zeng et al., 2025). Our taxonomy provides measurement, but the challenge lies in designing objectives and environments that induce these capabilities.

The Bidirectional Research Opportunity. The examples above illustrate how cognitive science provides principled frameworks for technical development. The relationship is also bidirectional: models provide unprecedented tools for testing cognitive theories at scale. Traditional cognitive research faces severe constraints in experimental control, sample size, and the ability to manipulate internal representations. Models overcome these limitations.

Models serve as computational implementations of cognitive hypotheses that can be systematically manipulated and tested (Griffiths et al., 2019; Saxe et al., 2021). We can examine how reasoning behaviors emerge during training (Chang & Bergen, 2024; Wei et al., 2022a), ablate architectural components to identify their functional roles (Sternberg et al., 2009), and test theories of hierarchical planning and meta-cognition through interventions impossible with human subjects. Recent work uses models to test theories of semantic cognition (Grand et al., 2022), conceptual knowledge (Patel et al., 2022), and pragmatic reasoning (Tessler & Goodman, 2016) at scales unachievable in traditional experiments.

Systematic differences between human and model reasoning provide constraints for cognitive theories. Our finding that humans deploy meta-cognitive monitoring while models do not, despite both succeeding on the same problems, suggests these behaviors enable generalization or error recovery beyond immediate task success. This hypothesis becomes testable through systematic manipulation of models. Recent work has revealed similar divergences in causal reasoning (Lampinen et al., 2022), social cognition (Sap et al., 2022), and pragmatic inference (Hu et al., 2023), each providing empirical constraints for theories of human cognition.

When models succeed through unexpected mechanisms, they challenge assumptions about necessary cognitive structures. Human reasoning patterns have already informed architectural innovations (Graves et al., 2016; Nye et al., 2021), while model capabilities have refined theories about what computations are sufficient for intelligent behavior. Our taxonomy enables this synergy by providing shared vocabulary. Cognitive science offers theories of what matters and why. Machine learning offers implementations to test at scale. Our framework provides the language connecting them, enabling not just better models but better theories tested and refined through computational implementation.

Acknowledgments

We wish to express our sincere gratitude to Anshul Nasery, Kshitish Ghate, Divyansh Pareek, Moe Kayali, Runjia Li, Harshita Chopra, Mihir Kavishwar, Ishika Agarwal, and Amruta Parulekar for their help in collecting human reasoning traces.

References

- Orevaoghene Ahia, Martijn Bartelds, Kabir Ahuja, Hila Gonen, Valentin Hofmann, Siddhant Arora, Shuyue Stella Li, Vishal Puttagunta, Mofetoluwa Adeyemi, Charishma Buchireddy, Ben Walls, Noah Bennett, Shinji Watanabe, Noah A. Smith, Yulia Tsvetkov, and Sachin Kumar. Blab: Brutally long audio bench, 2025. URL <https://arxiv.org/abs/2505.03054>.
- Jonathan Baron. *Thinking and deciding*. Cambridge University Press, 2008. URL <https://www.cambridge.org/core/books/thinking-and-deciding/8A7447BF6291D2FB365AF3B2F56FB9FB>.
- F. C. Bartlett. *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press, Cambridge, UK, 1932.
- Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.
- Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on "a is b" fail to learn "b is a". *arXiv preprint arXiv:2309.12288*, 2023.
- Matthew M Botvinick, Yael Niv, and Andrew G Barto. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *cognition*, 113(3):262–280, 2009.
- Robert Boyd, Peter J Richerson, and Joseph Henrich. The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, 108(supplement_2):10918–10925, 2011.
- Daniel A Braun, Carsten Mehring, and Daniel M Wolpert. Structure learning in action. *Behavioural brain research*, 206(2):157–165, 2010.
- Bruce K Britton and Abraham Tesser. Effects of prior knowledge on use of cognitive capacity in three complex cognitive tasks. *Journal of verbal learning and verbal behavior*, 21(4):421–436, 1982.
- Donald E. Broadbent. *Perception and Communication*. Pergamon Press, London, 1958.
- Tyler A Chang and Benjamin K Bergen. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350, 2024.
- Eugene Charniak. *Introduction to artificial intelligence*. Pearson Education India, 1985.
- William G Chase and Herbert A Simon. Perception in chess. *Cognitive psychology*, 4(1):55–81, 1973.
- Yuri Chervonyi, Trieu H Trinh, Miroslav Olšák, Xiaomeng Yang, Hoang Nguyen, Marcelo Menegali, June-hyuk Jung, Vikas Verma, Quoc V Le, and Thang Luong. Gold-medalist performance in solving olympiad geometry with alphageometry2. *arXiv preprint arXiv:2502.03544*, 2025.
- Michelene TH Chi, Paul J Feltovich, and Robert Glaser. Categorization and representation of physics problems by experts and novices. *Cognitive science*, 5(2):121–152, 1981.
- Noam Chomsky. *Aspects of the Theory of Syntax*. Number 11. MIT press, 2014.
- Allan M Collins and Elizabeth F Loftus. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407, 1975.
- Cathryn S Cortesa, Jonathan D Jones, Gregory D Hager, Sanjeev Khudanpur, Amy L Shelton, and Barbara Landau. Characterizing spatial construction processes: Toward computational tools to understand cognition. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 39, 2017.
- Fiery Cushman and Adam Morris. Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences*, 112(45):13817–13822, 2015.

Payel Das, Ching-Yun Ko, Sihui Dai, Georgios Kollias, Subhajit Chaudhury, and Aurelie Lozano. Can memory-augmented language models generalize on reasoning-in-a-haystack tasks? *arXiv preprint arXiv:2503.07903*, 2025.

Randall Davis and Jonathan J King. The origin of rule-based systems in ai. *Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project*, 1984.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhushu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanja Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.

Ray J Dolan and Peter Dayan. Goals and habits in the brain. *Neuron*, 80(2):312–325, 2013.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36:70293–70332, 2023.

Hermann Ebbinghaus. *Über das gedächtnis: untersuchungen zur experimentellen psychologie*. Duncker & Humblot, 1885.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Jonathan St. B. T. Evans. In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10):454–459, 2003. doi: 10.1016/j.tics.2003.08.012. URL <https://doi.org/10.1016/j.tics.2003.08.012>.

Jonathan St BT Evans and Keith E Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241, 2013.

-
- Evernote. AI Transcribe by Evernote. URL <https://evernote.com/ai-transcribe>.
- Leon Festinger. Cognitive dissonance. *Scientific American*, 207(4):93–106, 1962.
- John H Flavell. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906, 1979.
- Stephen M Fleming. Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, 75(1):241–268, 2024.
- Stephen M Fleming and Nathaniel D Daw. Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological review*, 124(1):91, 2017.
- Jerry Fodor. Language, thought and compositionality. *Royal Institute of Philosophy Supplements*, 48:227–242, 2001.
- Jerry A Fodor. *The language of thought*, volume 5. Harvard university press, 1975.
- Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.
- Steven M Frankland and Joshua D Greene. Concepts and compositionality: in search of the brain’s language of thought. *Annual review of psychology*, 71(1):273–303, 2020.
- Chris D Frith and Uta Frith. Social cognition in humans. *Current biology*, 17(16):R724–R732, 2007.
- Eugene Galanter et al. Plans and the structure of behavior. 1960.
- Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.
- Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170, 1983.
- Mary L Gick and Keith J Holyoak. Analogical problem solving. *Cognitive psychology*, 12(3):306–355, 1980.
- Mary L Gick and Keith J Holyoak. Schema induction and analogical transfer. *Cognitive psychology*, 15(1):1–38, 1983.
- Michael H Goldstein, Heidi R Waterfall, Arnon Lotem, Joseph Y Halpern, Jennifer A Schwade, Luca Onnis, and Shimon Edelman. General cognitive principles for learning structure in time and space. *Trends in cognitive sciences*, 14(6):249–258, 2010.
- Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3, 2004.
- Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, 6(7):975–987, 2022.
- Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016.
- Thomas L Griffiths, Frederick Callaway, Michael B Chang, Erin Grant, Paul M Krueger, and Falk Lieder. Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, 29:24–30, 2019.
- Thomas L Griffiths, Nick Chater, and Joshua B Tenenbaum. *Bayesian models of cognition: Reverse engineering the mind*. MIT Press, 2024.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanja Zhao, John Yang, Shreyas Pimpalaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models, 2025. URL <https://arxiv.org/abs/2506.04178>.

Graeme S Halford. Cognitive processing capacity and learning ability: An integration of two areas. *Learning and Individual Differences*, 1(1):125–153, 1989.

Eddie Harmon-Jones and Judson Mills. An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. 2019.

Grietjie Haupt. Hierarchical thinking: a cognitive tool for guiding coherent decision making in design problem solving. *International Journal of Technology and Design Education*, 28(1):207–237, 2018.

Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. The faculty of language: what is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579, 2002.

Fritz Heider. *The Psychology of Interpersonal Relations*. Wiley, New York, 1958.

Christoph Hoerl and Teresa McCormack. Thinking in and about time: A dual systems perspective on temporal cognition. *Behavioral and Brain Sciences*, 42:e244, 2019.

Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A Smith. Fluid language model benchmarking. *arXiv preprint arXiv:2509.11106*, 2025.

Keith J Holyoak and Paul Thagard. Analogical mapping by constraint satisfaction. *Cognitive science*, 13(3):295–355, 1989.

Keith J Holyoak and Paul Thagard. The analogical mind. *American psychologist*, 52(1):35, 1997.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4194–4213, 2023.

Clark L Hull. Quantitative aspects of evolution of concepts: An experimental study. *Psychological monographs*, 28(1):i, 1920.

Quentin JM Huys, Neir Eshel, Elizabeth O’Nions, Luke Sheridan, Peter Dayan, and Jonathan P Roiser. Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, 8(3):e1002410, 2012.

Robert A Jacobs and John K Kruschke. Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(1):8–21, 2011.

William James, Frederick Burkhardt, Fredson Bowers, and Kestutis Skrupskelis. *The principles of psychology*, volume 1. Macmillan London, 1890.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*, 2024.

Philip N Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010. URL <https://www.pnas.org/content/107/43/18243>.

Philip Nicholas Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press, 1983.

David H Jonassen. Toward a design theory of problem solving. *Educational technology research and development*, 48(4):63–85, 2000.

David H Jonassen and Woei Hung. All problems are not equal: Implications for problem-based learning. *Essential readings in problem-based learning: Exploring and extending the legacy of Howard S. Barrows*, 1741, 2015.

Priyanka Kargupta, Runchu Tian, and Jiawei Han. Beyond true or false: Retrieval-augmented hierarchical analysis of nuanced claims. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29664–29679, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1434. URL <https://aclanthology.org/2025.acl-long.1434/>.

Sangeet S Khemlani, Aron K Barbey, and Philip N Johnson-Laird. Causal reasoning with mental models. *Frontiers in human neuroscience*, 8:849, 2014.

Marina A Kholodnaya and Elena V Volkova. Conceptual structures, conceptual abilities and productivity of cognitive functioning: the ontological approach. *Procedia-Social and Behavioral Sciences*, 217:914–922, 2016.

Günther Knoblich, Stellan Ohlsson, Hilde Haider, and Detlef Rhenius. Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, memory, and cognition*, 25(6):1534, 1999.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 36, 2023. URL <https://arxiv.org/abs/2205.11916>.

Peter M Krafft and Thomas L Griffiths. Levels of analysis in computational social science. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 40, 2018.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenqi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, William Fineas Ellingwood, Sathvik Udupa, Siyuan Hou, Allison Ferner, Sara Barahona, Cecilia Bolaños, Satish Rahi, Laura Herrera-Alarcón, Satvik Dixit, Siddhi Patil, Soham Deshmukh, Lasha Koroshinadze, Yao Liu, Leibny Paola Garcia Perera, Eleni Zanou, Themos Stafylakis, Joon Son Chung, David Harwath, Chao Zhang, Dinesh Manocha, Alicia Lozano-Diez, Santosh Kesiraju, Sreyan Ghosh, and Ramani Duraiswami. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv preprint arXiv:2508.13992*, 2025. URL <https://arxiv.org/abs/2508.13992>.

Wolfgang Köhler. *The Mentality of Apes*. K. Paul, Trench, Trubner & co., ltd., London, 1925.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024. URL <https://arxiv.org/abs/2411.15124>.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 537–563, 2022.

Barbara Landau and Ray Jackendoff. Whence and whither in spatial language and spatial cognition? *Behavioral and brain sciences*, 16(2):255–265, 1993.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

-
- Karl Spencer Lashley et al. *The problem of serial order in behavior*, volume 21. Bobbs-Merrill Oxford, 1951.
- Hung Le, Dai Do, Dung Nguyen, and Svetha Venkatesh. Reasoning under 1 billion: Memory-augmented reinforcement learning for large language models. *arXiv preprint arXiv:2504.02273*, 2025.
- Mark R Leary and Nicole R Buttermore. The evolution of the human self: Tracing the natural history of self-awareness. *Journal for the theory of Social Behaviour*, 33(4):365–404, 2003.
- Seungpil Lee, Woochang Sim, Donghyeon Shin, Wongyu Seo, Jiwon Park, Seokki Lee, Sanha Hwang, Sejin Kim, and Sundong Kim. Reasoning abilities of large language models: In-depth analysis on the abstraction and reasoning corpus. *ACM Transactions on Intelligent Systems and Technology*, 2024.
- Yanyu Lei. Sociality and self-awareness in animals. *Frontiers in Psychology*, 13:1065638, 2023.
- Ang Li, Charles Wang, Kaiyu Yue, Zikui Cai, Ollie Liu, Deqing Fu, Peng Guo, Wang Bill Zhu, Vatsal Sharan, Robin Jia, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. Zebra-cot: A dataset for interleaved vision language reasoning, 2025a. URL <https://arxiv.org/abs/2507.16746>.
- Shuyue Stella Li, Avinandan Bose, Faeze Brahman, Simon Shaolei Du, Pang Wei Koh, Maryam Fazel, and Yulia Tsvetkov. Personalized reasoning: Just-in-time personalization and why llms fail at it. *arXiv preprint arXiv:2510.00177*, 2025b.
- Shuyue Stella Li, Melanie Sclar, Hunter Lang, Ansong Ni, Jacqueline He, Puxin Xu, Andrew Cohen, Chan Young Park, Yulia Tsvetkov, and Asli Celikyilmaz. Prefpalette: Personalized preference modeling with latent attributes, 2025c. URL <https://arxiv.org/abs/2507.13541>.
- Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024.
- Falk Lieder and Thomas L Griffiths. Strategy selection as rational metareasoning. *Psychological review*, 124(6):762, 2017.
- Falk Lieder and Thomas L Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43:e1, 2020.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. URL <https://arxiv.org/abs/2305.20050>.
- Tania Lombrozo. Learning by thinking in natural and artificial minds. *Trends in Cognitive Sciences*, 2024.
- Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. <https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2>, 2025. Notion Blog.
- Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, Tianrui Wang, Yuping Wang, Yuxuan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou, Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, Eng-Siong Chng, and Xie Chen. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*, 2025. URL <https://arxiv.org/abs/2505.13032>.
- Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, et al. Deepseek-r1 thoughtology: Let’s think about llm reasoning. *arXiv preprint arXiv:2504.07128*, 2025.

David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, San Francisco, 1982. ISBN 0-7167-1284-9.

Rui Mata, Bettina von Helversen, and Jörg Rieskamp. When easy comes hard: The development of adaptive strategy selection. *Child Development*, 82(2):687–700, 2011.

James L McClelland, Matthew M Botvinick, David C Noelle, David C Plaut, Timothy T Rogers, Mark S Seidenberg, and Linda B Smith. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, 14(8):348–356, 2010.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D Hardy, and Thomas L Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121, 2024.

Douglas L Medin and Marguerite M Schaffer. Context theory of classification learning. *Psychological review*, 85(3):207, 1978.

Smitha Milli, Falk Lieder, and Thomas L Griffiths. A rational reinterpretation of dual-process theories. *Cognition*, 217:104881, 2021.

Marvin Minsky. A framework for representing knowledge. 1974.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025. URL <https://arxiv.org/abs/2501.19393>.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Ulric Neisser. Five kinds of self-knowledge. *Philosophical psychology*, 1(1):35–59, 1988.

Thomas O Nelson. Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation*, volume 26, pp. 125–173. Elsevier, 1990.

Allen Newell and Herbert A. Simon. *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ, 1972. ISBN 978-0-13-445403-0. URL <https://www.pearson.com/us/higher-education/program/Newell-Human-Problem-Solving/PGM332539.html>.

Allen Newell, John C Shaw, and Herbert A Simon. Report on a general problem solving program. In *IFIP congress*, volume 256, pp. 1959. Pittsburgh, PA, 1959.

Nils J. Nilsson. Problem-solving methods in artificial intelligence. In *McGraw-Hill computer science series*, 1971. URL <https://api.semanticscholar.org/CorpusID:34428834>.

Laura R Novick and Cindy E Hmelo. Transferring symbolic representations across nonisomorphic problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(6):1296, 1994.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. 2021.

Mike Oaksford and Nick Chater. Précis of bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences*, 32(1):69–84, 2009.

Stellan Ohlsson. Information-processing explanations of insight and related phenomena. *Advances in the psychology of thinking*, pp. 1–44, 1992.

Olmo Team, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lj Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3 technical report. 2025.

Jooyoung Park, Fang-Chi Lu, and William M Hedgcock. Relative effects of forward and backward planning on goal pursuit. *Psychological science*, 28(11):1620–1630, 2017.

Ajay Patel, Bryan Li, Mohammad Sadegh Rasooli, Noah Constant, Colin Raffel, and Chris Callison-Burch. Bidirectional language models are also few-shot learners. *arXiv preprint arXiv:2209.14500*, 2022.

David Peebles and Richard P Cooper. Thirty years after marr’s vision: levels of analysis in cognitive science, 2015.

John Piaget. The origins of intelligence in children. *International Universities*, 1952.

Steven Thomas Piantadosi. *Learning and the language of thought*. PhD thesis, Massachusetts Institute of Technology, 2011.

Michael I Posner and Steven W Keele. On the genesis of abstract ideas. *Journal of experimental psychology*, 77(3p1):353, 1968.

Tian Qin, David Alvarez-Melis, Samy Jelassi, and Eran Malach. To backtrack or not to backtrack: When sequential search limits model reasoning. *arXiv preprint arXiv:2504.07052*, 2025.

M Ross Quillan. Semantic memory. Technical report, 1966.

Marco Ragni, Ilir Kola, and Phil N Johnson-Laird. The wason selection task: A meta-analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 39, 2017.

Lance J Rips. Cognitive processes in propositional reasoning. *Psychol. Rev.*, 90(1):38–71, January 1983.

Bethany Rittle-Johnson, Robert S Siegler, and Martha Wagner Alibali. Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of educational psychology*, 93(2):346, 2001.

Philippe Rochat. Five levels of self-awareness as they unfold early in life. *Consciousness and cognition*, 12(4):717–731, 2003.

Eleanor Rosch. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192, 1975.

Eleanor Rosch. Principles of categorization. In Eleanor Rosch and Barbara Bloom Lloyd (eds.), *Cognition and Categorization*, pp. 27–48. Lawrence Elbaum Associates, 1978.

Eleanor Rosch and Carolyn B Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605, 1975.

David A Rosenbaum, Rajal G Cohen, Steven A Jax, Daniel J Weiss, and Robrecht Van Der Wel. The problem of serial order in behavior: Lashley’s legacy. *Human movement science*, 26(4):525–554, 2007.

Jacob Russin, Sam Whitman McGrath, and Danielle J Williams. From frege to chatgpt: Compositionality in language, cognition, and deep neural networks. *arXiv preprint arXiv:2405.15164*, 2024.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 3762–3780, 2022.

Andrew Saxe, Stephanie Nelli, and Christopher Summerfield. If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, 22(1):55–67, 2021.

Alan H Schoenfeld. *Mathematical problem solving*. Elsevier, 2014.

O. G. Selfridge. Pandemonium: A paradigm for learning. In *Proceedings of the Symposium on Mechanisation of Thought Processes*, pp. 513–526, London, 1959. Her Majesty’s Stationery Office.

Patrick Shafto, Charles Kemp, Elizabeth Baraff Bonawitz, John D Coley, and Joshua B Tenenbaum. Inductive reasoning about causally transmitted properties. *Cognition*, 109(2):175–192, 2008.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, et al. Spurious rewards: Rethinking training signals in rlvr. *arXiv preprint arXiv:2506.10947*, 2025.

Amy Lynne Shelton, E Emory Davis, Cathryn S Cortesa, Jonathan D Jones, Gregory D Hager, Sanjeev Khudanpur, and Barbara Landau. Characterizing the details of spatial construction: cognitive constraints and variability. *Cognitive Science*, 46(1):e13081, 2022.

Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022.

Dorothea P. Simon and Herbert A. Simon. Individual differences in solving physics problems. In *Children’s Thinking: What Develops?*, pp. 24, Hillsdale, N.J., 1978. Lawrence Erlbaum Associates.

B. F. Skinner. *Science and Human Behavior*. Simon & Schuster, New York, 1953.

Steven Sloman and Steven A Sloman. *Causal models: How people think about the world and its alternatives*. Oxford University Press, 2009.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Robert J Sternberg, Karin Sternberg, and Jeff Mio. *Cognitive psychology*. wadsworth Belmont, CA, 2009.

Stanley Smith Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946.

Mark Steyvers and Joshua B Tenenbaum. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1):41–78, 2005.

Deborah Stipek, Susan Recchia, Susan McClintic, and Michael Lewis. Self-evaluation in young children. *Monographs of the society for research in child development*, pp. i–95, 1992.

John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285, 1988.

John Sweller. Cognitive load theory. In *Psychology of learning and motivation*, volume 55, pp. 37–76. Elsevier, 2011.

Ross Taylor. Generalthoughtarchive. <https://huggingface.co/datasets/RJT1990/GeneralThoughtArchive>, 2024. Dataset on Hugging Face; MIT License.

Teknium, Roger Jin, Chen Guang, Jai Suphavadeeprasit, and Jeffrey Quesnelle. Deephermes 3 preview, 2025.

-
- Michael Henry Tessler and Noah D Goodman. A pragmatic theory of generic language. *arXiv preprint arXiv:1608.02926*, 2016.
- Paul Thagard. *Coherence in thought and action*. MIT press, 2002.
- Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.
- Anne Treisman. Monitoring and storage of irrelevant messages in selective attention. *Journal of Verbal Learning and Verbal Behavior*, 3(6):449–459, 1964.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.
- Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185 (4157):1124–1131, 1974. doi: 10.1126/science.185.4157.1124. URL <https://doi.org/10.1126/science.185.4157.1124>.
- Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- Peter C Wason. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273–281, 1968.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022b. URL <https://arxiv.org/abs/2201.11903>.
- Max Wertheimer. *Productive Thinking*. Harper and Brothers, New York, 1945.
- Robert A Wicklund. The influence of self-awareness on human behavior: The person who becomes self-aware is more likely to act consistently, be faithful to societal norms, and give accurate reports about himself. *American scientist*, 67(2):187–193, 1979.
- Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Huijie Lv, Ming Zhang, et al. Reasoning or memorization? unreliable results of reinforcement learning due to data contamination. *arXiv preprint arXiv:2507.10532*, 2025.
- Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. Are large language models really good logical reasoners? a comprehensive evaluation and beyond. *IEEE Transactions on Knowledge and Data Engineering*, 2025.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yingger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2023. URL <https://arxiv.org/abs/2305.10601>.

Nick Yeung and Christopher Summerfield. Metacognition in human decision-making: confidence and error monitoring. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 367(1594):1310–1321, May 2012.

Zhiyuan Zeng, Hamish Ivison, Yiping Wang, Lifan Yuan, Shuyue Stella Li, Zhuorui Ye, Siting Li, Jacqueline He, Runlong Zhou, Tong Chen, et al. Rlve: Scaling up reinforcement learning for language models with adaptive verifiable environments. *arXiv preprint arXiv:2511.07317*, 2025.

A Appendix

A.1 Prompts for Fine-Grained Cognitive Element Annotation

We provide detailed annotation guidelines for each cognitive capability to ensure consistent and reliable human annotations. Below we present the complete annotation guidelines for *abstraction* as an illustrative example. The guidelines include: (1) a clear definition of the cognitive capability, (2) specific indicators to look for in reasoning traces, (3) a three-level rubric (0=Absent, 1=Partially Present, 2=Present), and (4) annotated examples demonstrating each score level. Complete annotation prompts for all cognitive capabilities are available in our repository at <https://github.com/stellalisy/CognitiveFoundations>.

Annotation Guidelines: Abstraction in the Reasoning Process

Definition: *Abstraction* is the ability to extract general principles from specific instances. In reasoning traces, abstraction refers to when the participant demonstrates the ability to identify underlying concepts, generalize from concrete examples, derive broader principles, and apply general concepts across different contexts.

What to Look For:

When analyzing a reasoning trace, look for evidence that the participant demonstrates abstraction:

1. **Generalization from examples:** Does the participant derive general principles from specific instances?
 - Look for extraction of broader patterns or rules from concrete cases
 - Check if the participant identifies commonalities that transcend specific examples
2. **Concept formation:** Does the participant form abstract concepts beyond surface features?
 - Look for formulation of higher-level constructs or categories
 - Check if the participant develops conceptual frameworks that organize specific instances
3. **Level shifting:** Does the participant move between concrete and abstract levels?
 - Look for transitions between specific examples and general principles
 - Check if the participant can apply abstract ideas to specific cases and extract abstractions from specifics
4. **Cross-domain application:** Does the participant apply principles across different domains?
 - Look for transfer of abstract concepts between distinct contexts
 - Check if the participant recognizes when the same abstract principle applies in different situations

Label Levels:

0 - Absent: The reasoning trace shows little to no abstraction. The participant focuses on specific details or concrete examples without extracting general principles or forming abstract concepts.

1 - Partially Present: The reasoning trace shows some abstraction, but with limited depth or inconsistent application. The participant occasionally generalizes from examples or forms basic abstractions, but doesn't consistently operate at an abstract level or effectively move between concrete and abstract.

2 - Present: The reasoning trace shows clear abstraction throughout. The participant consistently generalizes from specific instances, forms sophisticated abstract concepts, effectively moves between concrete and abstract levels, and applies principles across different domains.

Output Format:

First, write **###EXPLANATION** on its own line, followed by a brief one-sentence explanation of your reasoning about whether abstraction is present in the reasoning trace on the next line. Then write **###SCORE** on its own line, followed by your final score (0–2) on the next line.

The guidelines also include three detailed annotated examples demonstrating scores of 0 (Absent), 1 (Partially Present), and 2 (Present), which help annotators calibrate their judgments. Complete examples and guidelines for all other cognitive capabilities follow a similar structure and can be found in the repository.

Below we provide the full prompt shown to annotators for this capability.

A.2 Typology of Problems

We classify cognitive tasks using an extended version of Jonassen's (2000) problem-solving taxonomy. Jonassen's framework characterizes problems along a continuum from well-structured (clear goals, known solution paths, predictable outcomes) to ill-structured (ambiguous goals, multiple solution paths, uncertain outcomes), organizing problems into 11 types based on their structural properties and cognitive demands.

A.2.1 Extension of Jonassen's Taxonomy

We extend the original 11-category framework with two additional categories to capture tasks outside the goal-directed transformation paradigm:

- **Factual Recall:** Retrieving stored knowledge without requiring reasoning or problem-solving (e.g., "What is photosynthesis?" or "List the causes of WWI")
- **Creative/Expressive:** Generating novel content judged by originality or aesthetic quality rather than convergence to a predetermined solution (e.g., "Write a poem" or "Draw how you feel")

This yields a 13-category taxonomy spanning the full spectrum of cognitive demands in our datasets.

A.2.2 Problem Type Definitions

Following Jonassen (2000), we define problem-solving as a goal-directed cognitive activity that transforms an initial state into a desired goal state through systematic reasoning. The 11 problem-solving types are organized along the structuredness continuum:

Well-Structured Problems

1. **Logical:** Abstract reasoning puzzles with optimal solutions and minimal context (e.g., Tower of Hanoi, river crossing puzzles)
2. **Algorithmic:** Fixed procedures applied to similar variable sets, producing correct answers through prescribed methods (e.g., solve quadratic equations, convert temperature units)
3. **Story Problems:** Mathematical or scientific problems embedded in narrative contexts, requiring extraction of values and formula application (e.g., distance-rate-time problems)
4. **Rule-Using:** Procedural processes constrained by rules that allow multiple valid approaches to system-constrained answers (e.g., database searching, theorem proving, recipe modification)
5. **Decision-Making:** Selecting and justifying one option from a finite set of alternatives, weighing benefits and limitations (e.g., college selection, route planning, benefits package choice)

Moderately Structured Problems

6. **Troubleshooting:** Diagnosing faults in malfunctioning systems by generating and testing hypotheses (e.g., car won't start, network is down, code debugging)
7. **Diagnosis-Solution:** Extending beyond fault identification to recommend and evaluate treatment options (e.g., medical diagnosis and treatment, identifying and treating lawn problems)
8. **Strategic Performance:** Real-time execution of complex tactics while maintaining situational awareness under competing demands (e.g., flying aircraft, teaching live classes, managing portfolios during trading)

Ill-Structured Problems

9. **Case Analysis:** Analyzing complex scenarios with multiple stakeholders and perspectives, arguing positions in detail-rich situations with ill-defined goals (e.g., business cases, legal judgments, policy recommendations)

-
10. **Design:** Creating new artifacts or systems that satisfy functional requirements, with solutions evaluated as better or worse rather than correct or incorrect (e.g., bridge design, curriculum development, marketing campaigns)
 11. **Dilemma:** Reconciling contradictory positions with no satisfactory solution that serves all perspectives (e.g., abortion policy, international conflicts, wealth redistribution)

A.2.3 Classification Methodology

Each problem is classified through majority voting across three frontier language models (GPT-4o-mini, Gemini-2.5-Pro, and Claude-Sonnet-4.5). Each model independently classifies the problem using detailed annotation guidelines based on Jonassen's (2000) structural criteria: goal clarity, solution determinacy, and domain constraints. Three-way disagreements occur in under 3% of cases and are adjudicated manually using these structural criteria. The complete classification prompt and guidelines are available in our code repository at <https://github.com/stellalisy/CognitiveFoundations>.

A.2.4 Key Distinctions

Several problem types share surface similarities but differ fundamentally in their cognitive demands:

- **Troubleshooting vs. Diagnosis-Solution:** Troubleshooting identifies faults; diagnosis-solution both identifies and treats
- **Story Problem vs. Factual Recall:** Story problems require calculation despite narrative framing; factual recall simply explains information
- **Decision-Making vs. Dilemma:** Decision-making has acceptable solutions; dilemmas have no satisfactory resolution for all parties
- **Design vs. Creative/Expressive:** Design has functional requirements and constraints; creative/expressive tasks involve pure expression
- **Algorithmic vs. Design:** Algorithms have one correct procedure; design problems have multiple valid approaches with better/worse solutions

This taxonomy enables systematic analysis of how problem structure affects behavioral manifestation and reasoning success across our dataset of 171,485 traces.

A.3 Accuracy Analysis

Table 5 presents the complete accuracy results for all 14 text reasoning models across the 13 problem types in our taxonomy. Numbers in parentheses indicate the number of problems evaluated for each model-type pair. The models span five major architectural families: Qwen3 (Alibaba's native thinking mode integration), DeepSeek-R1 and its distillations (knowledge transfer from 671B teacher), OpenThinker (data-quality focused), DeepScaleR (efficient RL), s1.1 (Qwen-based efficient training), and DeepHermes-3 (hybrid reasoning with user-controlled depth).

A.3.1 Key Observations

Performance by Problem Structure Accuracy patterns reveal strong relationships with problem structuredness following Jonassen's (2000) taxonomy. Well-structured problems show higher average accuracy: Story Problems (79.5%), Algorithmic (63.8%), and Factual Recall (61.8%). Moderately-structured problems show intermediate performance: Troubleshooting (54.6%), Rule-Using (54.4%), Decision-Making (55.7%), Logical (56.4%). Ill-structured problems show lower accuracy: Diagnosis-Solution (44.7%), Design (46.6%), Case Analysis (53.5%). The notable exception is Dilemma (82.4%), which achieves high accuracy despite being the most ill-structured problem type—suggesting models excel at articulating positions even when no objectively correct solution exists.

Table 5: Accuracy by problem type and model for text reasoning tasks (Part 1). Numbers in parentheses indicate sample size.

Problem Type	Qwen3-32B	Qwen3-14B	Qwen3-8B	R1-Qwen-32B	R1-Qwen-14B	R1-Qwen-7B	R1-Qwen-1.5B
Algorithmic	78.4% (6274)	77.4% (6275)	74.8% (6274)	70.1% (6271)	66.1% (6275)	62.1% (6227)	47.7% (6027)
Story Problem	86.7% (113)	92.0% (113)	87.6% (113)	85.0% (113)	85.0% (113)	74.3% (113)	66.1% (112)
Rule-Using	79.8% (504)	61.9% (504)	61.3% (504)	59.3% (504)	54.2% (504)	50.2% (490)	28.3% (488)
Decision-Making	70.7% (99)	76.8% (99)	75.8% (99)	57.6% (99)	55.6% (99)	38.4% (99)	23.2% (99)
Troubleshooting	82.4% (91)	62.6% (91)	65.9% (91)	67.0% (91)	57.1% (91)	18.7% (91)	11.0% (91)
Diagnosis-Solution	78.3% (83)	67.5% (83)	54.2% (83)	47.0% (83)	47.0% (83)	4.8% (83)	2.4% (83)
Case Analysis	84.3% (121)	82.6% (121)	74.4% (121)	47.9% (121)	58.7% (121)	18.2% (121)	2.5% (121)
Design	75.2% (157)	66.2% (157)	64.3% (157)	50.0% (156)	40.8% (157)	28.2% (156)	9.2% (153)
Dilemma	99.1% (1166)	97.6% (1168)	98.9% (1168)	94.5% (1167)	96.0% (1168)	89.7% (1168)	3.2% (1166)
Logical	73.3% (60)	78.3% (60)	68.3% (60)	73.3% (60)	68.3% (60)	48.3% (60)	18.2% (55)
Factual Recall	82.6% (2819)	80.2% (2819)	78.2% (2819)	68.0% (2819)	67.3% (2819)	42.0% (2818)	21.9% (2812)
Creative/Expressive	85.7% (7)	85.7% (7)	85.7% (7)	57.1% (7)	57.1% (7)	28.6% (7)	0.0% (7)
Mean	81.3%	77.7%	74.5%	64.7%	63.6%	41.9%	27.8%

Table 6: Accuracy by problem type and model (Part 2) and average across all models.

Problem Type	R1-Llama-70B	R1-Llama-8B	Hermes-8B	OpenThinker	DeepScaleR	s1.1-32B	R1-671B	Avg.
Algorithmic	68.3% (6276)	50.9% (6269)	31.0% (6276)	72.4% (6275)	50.6% (6195)	61.9% (6276)	81.6% (6276)	63.8%
Story Problem	83.2% (113)	73.5% (113)	57.5% (113)	85.0% (113)	77.3% (110)	72.6% (113)	87.6% (113)	79.5%
Rule-Using	57.9% (504)	36.1% (504)	32.9% (504)	75.0% (504)	30.5% (502)	49.0% (504)	85.7% (504)	54.4%
Decision-Making	64.6% (99)	48.5% (99)	40.4% (99)	74.7% (99)	14.1% (99)	57.6% (99)	81.2% (85)	55.7%
Troubleshooting	70.3% (91)	38.5% (91)	46.2% (91)	76.9% (91)	8.8% (91)	70.3% (91)	88.9% (90)	54.6%
Diagnosis-Solution	60.2% (83)	24.1% (83)	24.1% (83)	67.5% (83)	3.6% (83)	56.6% (83)	88.0% (83)	44.7%
Case Analysis	72.7% (121)	41.3% (121)	28.1% (121)	83.5% (121)	3.3% (121)	57.0% (121)	94.3% (106)	53.5%
Design	51.6% (157)	31.0% (155)	22.3% (157)	68.2% (157)	8.4% (155)	50.3% (157)	87.0% (131)	46.6%
Dilemma	95.6% (1168)	93.7% (1167)	89.6% (1168)	98.5% (1168)	3.4% (1168)	93.8% (1168)	100.0% (1)	82.4%
Logical	66.7% (60)	28.3% (60)	25.0% (60)	70.0% (60)	22.4% (58)	60.0% (60)	88.3% (60)	56.4%
Factual Recall	73.6% (2819)	47.3% (2819)	46.4% (2819)	80.2% (2819)	22.0% (2814)	67.8% (2819)	88.0% (2815)	61.8%
Creative/Expressive	57.1% (7)	57.1% (7)	42.9% (7)	57.1% (7)	14.3% (7)	42.9% (7)	85.7% (7)	54.1%
Mean	68.5%	47.5%	40.5%	75.8%	21.6%	61.7%	88.0%	59.0%

Frontier Model Performance **DeepSeek-R1-671B** (88.0% average) establishes the performance ceiling across nearly all problem types, achieving 81–88% on most categories. This 671B parameter MoE model (37B activated) underwent extensive multi-stage RL training, demonstrating the capabilities achievable with flagship-scale resources. Notably, R1-671B shows smallest gains over smaller models on Dilemma (100.0% with only 1 sample—unreliable) and Story Problems (87.6%), suggesting these problem types saturate more quickly with model capability. Largest improvements appear on ill-structured problems: Diagnosis-Solution (+43% over average), Case Analysis (+41%), and Design (+40%), confirming that complex multi-step reasoning benefits most from scale and sophisticated training.

Training Methodology Effects Performance patterns strongly reflect training approaches. The **Qwen3 series** (32B: 81.3%, 14B: 77.7%, 8B: 74.5%) achieves consistently strong performance approaching R1-671B’s frontier results, with the 32B variant achieving second-highest average accuracy across all models. This stems from comprehensive 4-stage RL training (cold-start SFT, reasoning RL with GRPO, thinking mode fusion, general RL) for the 32B flagship, followed by efficient strong-to-weak distillation for smaller variants. The smooth degradation (32B→14B: -3.6%, 14B→8B: -3.2%) demonstrates effective knowledge transfer through the two-phase distillation approach.

OpenThinker-32B (75.8% average) achieves third-highest performance despite using only 114K verified examples—86% less data than DeepSeek’s 800K distillation corpus. This model demonstrates that data quality trumps quantity, outperforming all R1 distillations except the 70B Llama variant. OpenThinker’s automated verification process (code execution for programming, LLM judge validation for mathematics) filters incorrect reasoning traces, producing cleaner training signals. Particularly strong on problems with objective correctness criteria: Story Problems (85.0%), Case Analysis (83.5%), Factual Recall (80.2%).

DeepSeek-R1 distillations show clear scaling effects. The Qwen-based variants (32B: 64.7%, 14B: 63.6%, 7B: 41.9%, 1.5B: 27.8%) demonstrate smooth performance degradation with model size under pure knowledge transfer via SFT on 800K teacher-generated examples. The Llama-based variants (70B: 68.5%, 8B: 47.5%) require substantially more parameters for equivalent performance—R1-Llama-70B needs 2.2× the parameters

of R1-Qwen-32B to achieve similar accuracy (68.5% vs 64.7%), confirming Qwen2.5’s superior parameter efficiency for reasoning tasks.

Problem Type Variability and Training Robustness Inter-model variance reveals which problem types expose training methodology differences. **Dilemma problems** show extreme variance (range: 3.2%–100.0%, SD ≈ 38%), with most models achieving high accuracy (90–99%) except the smallest distilled models (R1-Qwen-1.5B: 3.2%, DeepScaleR-1.5B: 3.4%). This suggests dilemma reasoning—requiring articulation and balancing of multiple contradictory positions—collapses catastrophically below critical capacity thresholds around 7–8B parameters. The 100% score for R1-671B reflects only a single evaluation sample and should not be interpreted as reliable.

Diagnosis-Solution problems exhibit the largest absolute variance (2.4%–88.0%, SD ≈ 28%), with severe degradation in smaller distilled models. R1-Qwen-7B drops precipitously to 4.8%, R1-Qwen-1.5B to 2.4%, and DeepScaleR-1.5B to 3.6%, while 30B+ models maintain 47–88% accuracy. This problem type requires coordinated fault identification and treatment evaluation—a multi-step reasoning process demanding sustained working memory and systematic hypothesis testing that smaller models cannot maintain coherently.

Story Problems show relatively consistent performance (range: 57.5%–92.0%, SD ≈ 10%), suggesting these well-structured problems are more robust to model capacity and training methodology differences. Even the smallest models (R1-Qwen-1.5B: 66.1%, DeepHermes-3-8B: 57.5%, DeepScaleR-1.5B: 77.3%) maintain reasonable performance.

Sample Size Considerations Problem type representation varies dramatically. **Algorithmic problems** dominate with 6,027–6,276 instances per model (59% of dataset). **Factual Recall** (2,812–2,819 instances, 27%) and **Dilemma** (1–1,168 instances, 11%) also show substantial representation for most models, though R1-671B’s single Dilemma sample renders its 100% score statistically meaningless. **Creative/Expressive** has only 7 examples, making results for this category unreliable.