Cognitive Foundations for Reasoning and Their Manifestation in LLMs

Anonymous authors
Paper under double-blind review

Abstract

Large language models (LLMs) solve complex problems yet fail on simpler variants, suggesting they achieve correct outputs through mechanisms fundamentally different from human reasoning. To understand this gap, we synthesize cognitive science research into a taxonomy of 28 cognitive elements spanning reasoning invariants, meta-cognitive controls, knowledge representations, and transformation operations. We conduct the first large-scale empirical analysis of 192K reasoning traces from 18 models across text, vision, and audio modalities, complemented by 54 human think-aloud traces. Our analysis reveals a fundamental misalignment: models narrow to rigid sequential processing on ill-structured problems precisely where diverse representations and meta-cognitive monitoring correlate most strongly with success. Human traces show more abstraction and conceptual processing, while models default to surface-level enumeration. Leveraging these behavioral patterns, we develop test-time reasoning guidance that scaffolds successful cognitive structures, improving performance by up to 26.7% on complex problems. This confirms that models possess latent reasoning capabilities but fail to deploy them spontaneously. Our framework establishes shared vocabulary between cognitive science and LLM research, enabling systematic diagnosis of reasoning failures and principled development of models that reason through robust cognitive mechanisms rather than spurious shortcuts.

1 Introduction

Humans extrapolate from existing knowledge to unfamiliar scenarios—a process that constitutes reasoning (Lombrozo, 2024). Large language models (LLMs), by contrast, exhibit failures of generalization that are unintuitive from the standpoint of human cognition: they master olympiad-level geometry (Chervonyi et al., 2025) while failing on simple cipher variants (McCoy et al., 2024; Mancoridis et al., 2025). This brittleness suggests models may achieve correct outputs through mechanisms fundamentally different from the robust cognitive structures that enable even young children to flexibly deploy novel concepts across unfamiliar domains (Bloom & Markson, 1998).

Understanding *why* LLMs fail to generalize is challenging because current paradigms reward reasoning outcomes without examining the cognitive processes that produce them (Lambert et al., 2025), conflating genuine reasoning with memorization (Wu et al., 2025). This creates a measurement crisis: we lack both the conceptual framework to

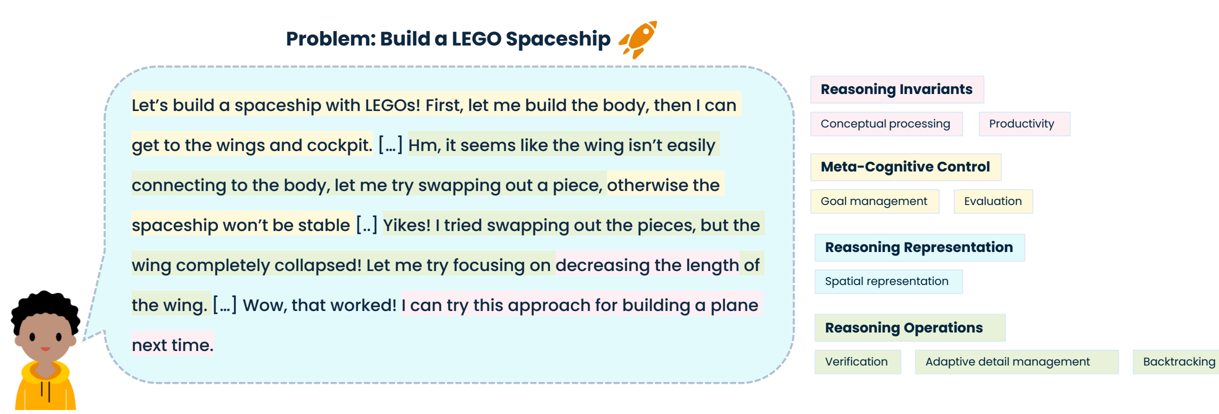

Figure 1: Cognitive elements in a reasoning trace for building a LEGO spaceship, characterized along the four dimensions of our taxonomy (Table 1).

characterize what cognitive elements should manifest in models and the empirical methods to assess whether they do. Contemporary research investigates scattered behaviors, such as verification (Gandhi et al., 2025), decomposition (Xu et al., 2025), and self-monitoring (Marjanović et al., 2026), but without systematic integration. By contrast, human reasoning research encompasses diverse phenomena including analogical transfer (Gentner, 1983), causal inference (Sloman & Sloman, 2009), representational flexibility (Ohlsson, 1992), and meta-cognitive control (Nelson, 1990). Without a comprehensive framework spanning these elements, we risk optimizing what we measure rather than what matters (Manheim & Garrabrant, 2018).

To address this gap, we introduce a **unified taxonomy of cognitive foundations** synthesizing established theories of problem-solving, mental representation, and meta-cognition (Table 1). Organized through Marr's (1982) levels of analysis, our taxonomy spans four dimensions: (1) *reasoning invariants*—computational-level properties that valid reasoning must maintain, such as logical coherence and compositionality; (2) *meta-cognitive controls*—executive functions that select and monitor reasoning strategies; (3) *reasoning representations*—formats for organizing knowledge hierarchically, causally, or spatially; and (4) *reasoning operations*—procedures that construct and transform representations, from verification to abstraction to backtracking. These 28 elements provide shared vocabulary for diagnosing which aspects of human reasoning manifest in LLMs.

Using this taxonomy, we conduct the **first large-scale empirical comparison of cognitive elements in human versus LLM reasoning**. We analyze 192K traces from 18 models across text, vision, and audio modalities, plus 54 human think-aloud traces, using fine-grained span-level annotation to identify which elements appear, where, and in what sequence. Our central finding reveals a fundamental misalignment: **models deploy cognitive elements inversely to what success requires**. On well-structured problems, models deploy diverse behavioral repertoires; as problems become ill-structured, they narrow to rigid sequential processing. This is precisely the opposite of successful traces, which show elevated meta-cognitive monitoring and representational diversity on complex problems.

We further introduce a **reasoning structure representation** encoding the temporal and hierarchical organization of cognitive elements within traces. Comparing frequent versus successful structures reveals systematic divergence: common reasoning patterns include elements with negative correlation to success, while successful traces exhibit deliberate problem scoping strategies that models bypass. Human traces show more abstraction and conceptual processing; models default to shallow forward chaining with limited error correction.

Leveraging these patterns, we develop **test-time reasoning guidance** that scaffolds successful cognitive structures. This intervention improves performance by up to 26.7% on ill-structured problems while maintaining baseline performance on well-structured ones, confirming that models possess latent capabilities they fail to deploy spontaneously. Our contributions include: (1) a cognitively-grounded taxonomy bridging human reasoning research and LLM evaluation; (2) the first large-scale empirical analysis revealing misalignment between deployed and success-correlated cognitive elements; (3) reasoning structure representations exposing systematic divergence from successful patterns; and (4) test-time guidance demonstrating these insights are actionable.

2 A Taxonomy of Cognitive Foundations

What cognitive capacities must a reasoner possess to solve problems effectively? We synthesize decades of cognitive science research to identify four functional requirements: reasoning must maintain validity constraints, deploy strategic control, organize knowledge in task-appropriate formats, and execute transformations on those structures. These requirements map onto Marr's (1982) levels of analysis: the first two specify *computational-level* goals (what reasoning must achieve), while the latter two specify *algorithmic-level* realizations (how reasoning proceeds). Table 1 operationalizes these dimensions into 28 cognitive elements grounded in foundational cognitive science research.

Maintaining Valid Inference: Reasoning Invariants. Reasoning must preserve properties that distinguish valid inference from arbitrary symbol manipulation. Fodor's (1975) Language of Thought Hypothesis identifies these properties: valid reasoning manipulates structured representations according to compositional rules, producing thoughts that are logically coherent, compositional, and productive (Fodor & Pylyshyn, 1988; Quilty-Dunn et al., 2023). *Logical coherence* maintains consistency across reasoning steps and contexts (Rips, 1983; Thagard, 2002)—simultaneously believing contradictory propositions creates cognitive dissonance requiring resolution (Festinger, 1962; Harmon-Jones & Mills, 2019). *Compositionality* enables building complex ideas from simpler components through rule-governed combination (Fodor, 1975; 2001), drastically increasing learning efficiency and expressive power (Frankland & Greene, 2020; Piantadosi, 2011). *Productivity* extends compositionality by enabling generation of infinitely many novel thoughts from finite primitives, intrinsically linked to systematicity (Fodor & Pylyshyn, 1988; Hauser et al., 2002). *Conceptual processing* operates over abstract semantic relations rather than surface forms (Halford, 1989; Kholodnaya & Volkova, 2016), enabling generalization beyond shallow pattern matching.

Directing the Reasoning Process: Meta-Cognitive Controls. Meta-cognitive controls constitute the executive functions that select, monitor, and adapt reasoning processes (Fleming, 2024). *Self-awareness* assesses one's knowledge state, capabilities, and task solvability (Neisser, 1988; Wicklund, 1979; Leary & Buttermore, 2003)—what Rochat (2003) describes as "arguably the most fundamental issue in psychology." *Context awareness* perceives situational demands, environmental constraints, and other agents (Boyd et al., 2011; Frith & Frith, 2007; Lei, 2023), determining which strategies are appropriate (Milli et al., 2021). *Strategy selection* chooses approaches suited to task demands (Lieder & Griffiths, 2017; 2020). *Goal management* establishes, sequences, and dynamically adjusts objectives throughout reasoning (Griffiths et al., 2019; Dolan & Dayan, 2013; Cushman & Morris, 2015). *Evaluation* monitors progress, assesses solution quality, and triggers adaptation (Fleming & Daw, 2017; Stipek et al., 1992).

Structuring Knowledge: Reasoning Representations. The effectiveness of reasoning depends critically on how information is structured (Sweller, 1988; Britton & Tesser, 1982), tracing back to early associationism (James, 1890). Cognitive load theory demonstrates that poorly structured information overwhelms working memory, while well-structured information facilitates processing (Sweller, 2011). *Structural organization* specifies how elements connect (Kemp & Tenenbaum, 2008): *sequential organization* orders steps where sequence matters (Skinner, 1965); *hierarchical organization* nests concepts in parent-child relationships (Miller et al., 2017; Botvinick et al., 2009; Haupt, 2018); *network organization* links concepts through multiple relationship types (Quillan, 1966; Shafto et al., 2008). *Conceptual organization* structures meaning: *ordinal organization* arranges elements by rank (Stevens, 1946); *causal organization* connects elements through cause-effect relations (Heider, 1958; Khemlani et al., 2014); *temporal organization* orders by before-after relations (Zakay & Block, 1997; Hoerl & McCormack, 2019); *spatial organization* structures elements geometrically (Tolman, 1948; Shepard & Metzler, 1971; Landau & Jackendoff, 1993; Shelton et al., 2022; Cortesa et al., 2017).

Transforming Representations: Reasoning Operations. Reasoning unfolds through goal-directed procedures that construct, evaluate, modify, and navigate representations (McClelland et al., 2010; Johnson-Laird, 1983). *Selection* operations align structures to demands: *context alignment* chooses schemas fitting the problem (Gick & Holyoak, 1980; 1983; Gentner, 1983); *knowledge alignment* maps problems onto domain-specific schemas (Chase & Simon, 1973; Chi et al., 1981; Schoenfeld, 2014). *Evaluation* through *verification* checks intermediate inferences for consistency and plausibility (Flavell, 1979; Goldstein et al., 2010). *Modification* operations adapt representations: *selective attention* focuses on relevant features (Broadbent, 1958; Treisman, 1964); *adaptive detail management* adjusts granularity (Rosch, 1978); *decomposition and integration* divides problems and synthesizes solutions (Newell et al., 1959); *representational restructuring* reformulates for insight (Wertheimer, 1945; Kohler, 2018; Braun et al., 2010); *pattern recognition* detects recurring structures (Selfridge, 1988; Posner & Keele, 1968); *abstraction* generalizes from instances (Hull, 1920; Piaget, 1952). *Navigation* operations traverse problem spaces: *forward chaining* reasons from facts toward goals (Huys et al., 2012; Newell & Simon, 1972; Davis & King, 1984); *backward chaining* works from goals to prerequisites (Park et al., 2017; Charniak, 1985; Newell et al., 1959); *backtracking* revisits prior paths upon detecting errors (Qin et al., 2025; Nilsson, 1971).

Dimensional Interaction. These dimensions form an integrated system rather than independent modules (Griffiths et al., 2015). Invariants constrain which operation sequences constitute valid reasoning. Meta-cognitive controls select strategies determining which representations to construct and operations to deploy. Representations provide the substrate over which operations act. Current research tends to isolate individual elements—studying verification independently of representation selection, or decomposition separately from meta-cognitive monitoring. Our taxonomy captures their systematic interaction, enabling diagnosis of where reasoning breaks down: in the constraints it violates, the strategies it selects, the representations it constructs, or the operations it executes.

3 Analyzing Cognitive Elements in Reasoning Traces

3.1 Dataset Overview

We analyze 192,709 reasoning traces from 18 models spanning three modalities. For **text reasoning**, we evaluate 16 open-weight models across multiple architecture families: Qwen3 (8B, 14B, 32B) with hybrid thinking mode (Yang et al., 2025), DeepSeek-R1 (671B) and its distillations across Qwen2.5 (1.5B–32B) and Llama3 (8B, 70B) bases (DeepSeek-AI et al., 2025), Olmo-3 (7B, 32B) with thinking mode (Olmo et al., 2025), and community reasoning models including OpenThinker-32B (Guha et al., 2025), s1.1-32B (Muennighoff et al., 2025), DeepScaleR-1.5B (Luo et al., 2025), and DeepHermes-3-8B (Teknium et al., 2025). Text problems comprise 10,612 questions from `GeneralThought` (Taylor, 2024) and ClaimSpect (Kargupta et al., 2025), the latter supplementing dilemma-type problems underrepresented in verifiable benchmarks. For **audio reasoning**, we evaluate Qwen3-Omni-30B on 4,917 problems from BLAB (Ahia et al., 2025), MMAR (Ma et al., 2025), and MMAU-Pro (Kumar et al., 2025). For **image reasoning**, we analyze 18,000 curated traces from Zebra-CoT (Li et al., 2025). We complement model traces with 54

Table 1: Taxonomy of 28 cognitive elements across four dimensions. Reasoning invariants and meta-cognitive controls specify computational-level goals; representations and operations specify algorithmic-level realizations.

A. Reasoning Invariants: Properties that valid reasoning must maintain.		
`Logical coherence`	Maintain consistency across reasoning steps and contexts (Fodor & Pylyshyn, 1988).	
`Compositionality`	Build complex ideas from simpler components (Fodor, 1975).	
`Productivity`	Generate unbounded thoughts from finite primitives (Halford, 1989).	
`Conceptual processing`	Operate over abstract representations rather than surface forms (Halford, 1989).	
B. Meta-Cognitive Controls: Executive functions that direct and monitor reasoning.		
`Self-awareness`	Assess own knowledge state, capabilities, and task solvability (Wicklund, 1979).	
`Context awareness`	Perceive situational demands, constraints, and other agents (Frith & Frith, 2007).	
`Strategy selection`	Choose reasoning approaches suited to problem structure (Lieder & Griffiths, 2017).	
`Goal management`	Establish, maintain, and adjust goals throughout reasoning (Griffiths et al., 2019).	
`Evaluation`	Monitor progress and solution quality; trigger adaptation (Fleming & Daw, 2017).	
C. Reasoning Representations: Formats for organizing knowledge and reasoning steps.		
Structural Organization	`Sequential organization`	Order steps where sequence matters (Skinner, 1965).
	`Hierarchical organization`	Nest concepts in parent-child relationships (Miller et al., 2017).
	`Network organization`	Link concepts through multiple relationship types (Quillan, 1966).
Conceptual Organization	`Ordinal organization`	Arrange elements by relative order or rank (Stevens, 1946).
	`Causal organization`	Connect elements through cause-effect relations (Heider, 1958).
	`Temporal organization`	Order elements by before-after relations (Zakay & Block, 1997).
	`Spatial organization`	Structure elements by spatial relationships (Tolman, 1948).
D. Reasoning Operations: Procedures that construct and transform representations.		
Selection	`Context alignment`	Align representations to task demands (Gick & Holyoak, 1980).
	`Knowledge alignment`	Align to domain-specific structures (Chi et al., 1981).
Evaluation	`Verification`	Check steps against predetermined criteria (Flavell, 1979).
Modification	`Selective attention`	Focus on relevant details; filter noise (Broadbent, 1958).
	`Adaptive detail management`	Adjust granularity based on task demands (Rosch, 1978).
	`Decomposition & integration`	Divide problems; synthesize subsolutions (Newell et al., 1959).
	`Representational restructuring`	Reformulate problems for new insights (Wertheimer, 1945).
	`Pattern recognition`	Detect recurring structures across contexts (Selfridge, 1988).
	`Abstraction`	Generalize from specific cases (Hull, 1920).
Navigation	`Forward chaining`	Reason from known facts toward goals (Huys et al., 2012).
	`Backward chaining`	Work backward from goals to prerequisites (Park et al., 2017).
	`Backtracking`	Revisit and correct prior reasoning paths (Nilsson, 1971).

human think-aloud recordings from 18 participants solving matched problems using a verbal protocol later transcribed for analysis. Complete dataset and model specifications appear in Appendix A.1.

Problem Classification. We classify problems using Jonassen's (2000) taxonomy, which organizes problem types along a continuum from *well-structured* (clear goals, known solution paths, verifiable outcomes) to *ill-structured* (ambiguous goals, multiple valid approaches, judgment-dependent evaluation). Well-structured problems include algorithmic, story, and rule-using tasks; mid-structured includes decision-making, troubleshooting, and diagnosis-solution; ill-structured problems include case analysis, design, and dilemmas. We extend this taxonomy with logic, factual recall, and creative/expressive categories for tasks outside goal-directed , yielding 13 problem types (Appendix A.3). Classification uses majority voting across three frontier models (GPT-4o-mini, Gemini-2.5-Pro, Claude-Sonnet-4.5), with manual adjudication for the $< 3\%$ of three-way disagreements.

Annotation and Evaluation. Each reasoning trace is annotated at the span level for the presence of all 28 cognitive elements. For each element, we develop annotation guidelines grounded in cognitive science definitions, including operational criteria, behavioral indicators, three-level scoring rubrics (absent/partial/present), and curated examples. The protocol requires marking exact text boundaries, enabling precise localization of cognitive behaviors. Guidelines were iteratively refined using human annotator feedback until convergence; full-scale annotation was performed using GPT-4.1 (Appendix A.2). Response correctness is evaluated using AlpacaEval (Dubois et al., 2024) with GPT-4o as judge, using ground-truth answers for verifiable tasks and Claude-Sonnet-4.5 responses for non-verifiable tasks.

3.2 Models Deploy Cognitive Elements Inversely to Success Requirements

We investigate the relationship between cognitive elements that models frequently deploy versus those that correlate most strongly with success. Figure 2 presents behavioral presence rates across problem types (left) alongside positive

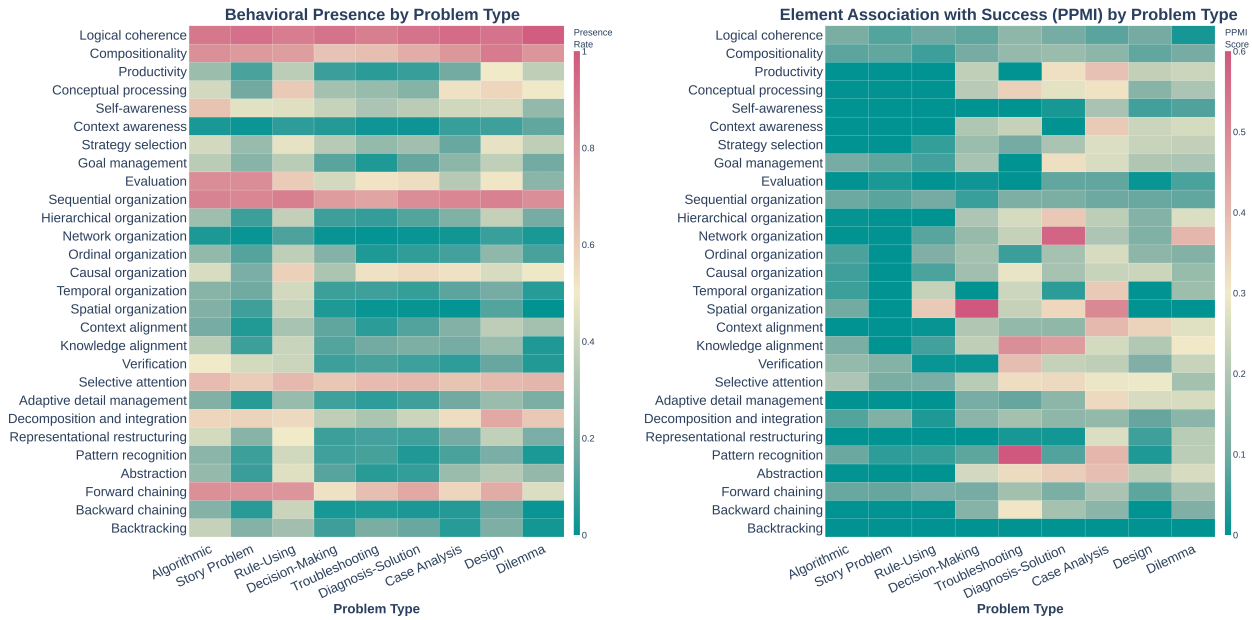

Figure 2: **Left:** Presence rate of each cognitive element by problem type, ordered from well-structured to ill-structured. **Right:** Positive pointwise mutual information (PPMI) between element presence and reasoning success. Models deploy broad behavioral repertoires on well-structured problems where success tolerates uniformity, but narrow to rigid sequential processing on ill-structured problems where success demands diversity.

pointwise mutual information (PPMI) between element presence and trace success (right). The contrast reveals a fundamental misalignment.

Models narrow their repertoire precisely where diversity matters most. On well-structured problems (algorithmic, story, rule-using), models deploy a broad repertoire of cognitive elements, with average presence of 0.397 ± 0.255 across all elements. As problems become ill-structured and non-verifiable, however, models *narrow* their behavioral deployment: case analysis, design, and dilemma problems show reduced average presence (0.337 ± 0.261), with usage concentrating heavily on sequential organization, logical coherence, and forward chaining.

This narrowing is precisely inverted from what success requires. The PPMI heatmap reveals that success on ill-structured problems demands *greater* behavioral diversity. Well-structured problems show low average PPMI (0.046 ± 0.063) across elements—many different behavioral combinations lead to success. Ill-structured problems show elevated PPMI across a wider range of elements (0.186 ± 0.114), particularly diverse representations (hierarchical, network, spatial, temporal) and varied operations (backward chaining, representational restructuring, pattern recognition). Models thus apply their richest cognitive repertoires where diversity matters least, while defaulting to limited, inflexible strategies precisely where adaptability is critical.

Frequent elements correlate weakly with success. Reasoning invariants, particularly logical coherence and compositionality, appear ubiquitously across problem types yet show surprisingly weak correlation with success. Logical coherence has average PPMI of 0.091 despite 91% presence rate; knowledge alignment shows PPMI of 0.234 while appearing in only 20.2% of traces. Manual inspection reveals a systematic pattern: while models frequently *attempt* to identify logical inconsistencies and contradictions, they consistently fail to recognize or effectively respond to them. This execution gap explains the discrepancy between high prevalence and low predictive value.

Meta-cognitive monitoring fails on ambiguous problems. Evaluation has a high presence (53.5%) but low success correlation (PPMI of 0.031), with frequency declining sharply for ill-structured problems. This pattern suggests models struggle particularly with self-assessment when ground truth is ambiguous or absent, precisely the conditions where meta-cognitive monitoring matters most.

Default representations hinder alternative organizational strategies. Models exhibit strong preference for sequential and causal organization regardless of problem type. However, success patterns indicate that as problems

Table 2: Illustrative examples of human versus LLM reasoning on matched problems. This table shows partial (for brevity) reasoning traces of both humans and DeepSeek-R1 (R1) on the same problems for both Logical and Design problems. Spans containing behavioral manifestation of cognitive elements, or lack thereof, are annotated as such: (***element***). Humans invoke abstraction and conceptual processing early, producing shorter traces; LLMs resort to surface-level enumeration and repetitive verification.

Type	Source	Reasoning Trace Excerpt
Logical	Human (166 tok.)	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 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***).
	R1 (7,632 tok.)	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 tok.)	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 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***).
	R1 (747 tok.)	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 ...

become less structured, diverse organizational strategies (hierarchical, network, ordinal, temporal, spatial) become increasingly critical. Spatial organization appears in only 9.8% of traces despite PPMI of 0.252. This reflects a fundamental challenge: when problems lack inherent structure, successful reasoners must *actively construct* appropriate organizational frameworks rather than defaulting to sequential processing.

3.3 Human Reasoners Deploy Different Cognitive Patterns

To understand what distinguishes successful reasoning, we compare cognitive element deployment between humans and LLMs on matched problems. Table 2 presents illustrative examples revealing systematic differences.

Humans invoke abstraction and conceptual processing early. On logical problems, humans quickly identify abstract principles, recognizing that diagonal corners share colors, that dominoes cover exactly one cell of each color. This led to dramatically shorter traces (166 vs. 7,632 tokens). LLMs instead resort to surface-level enumeration, explicitly counting individual squares rather than reasoning about structural properties. While LLMs eventually reach abstraction, they do so only after extensive concrete exploration.

Humans deploy richer meta-cognitive repertoires on open-ended problems. On design problems requiring the integration of multiple considerations, humans engage strategy selection, ordinal organization for ranking alternatives, context awareness for evaluating sources, and self-awareness when results are surprising. LLMs produce superficially reasonable responses but rely primarily on causal organization and decomposition, missing the meta-cognitive monitoring that enables humans to evaluate and refine their reasoning. This is reflect in the human exhibiting a longer trace than the LLM, the opposite observation from the logical trace lengths.

LLMs fail to learn from verification. Manual annotation reveals that while LLMs frequently attempt verification and backtracking, they fail to integrate feedback from these operations. Models often repeat verification on claims already explored and backtrack to paths already rejected. This pattern, attempting but failing to execute meta-cognitive monitoring, helps explain why these elements show high presence but low success correlation in Figure 2.

Implications for intervention. The cognitive structures humans naturally deploy (e.g., early abstraction, diverse representations, meta-cognitive monitoring) are precisely those which models under-utilize. If successful reasoning exhibits characteristic structural patterns beyond mere element presence, scaffolding models to follow these patterns may improve performance on ill-structured problems. This motivates our test-time reasoning guidance approach.

4 Scaffolding Successful Reasoning Structures

Our analysis reveals that models possess diverse cognitive capabilities but fail to deploy them in patterns associated with success. We investigate whether explicitly scaffolding successful reasoning structures can improve performance.

4.1 Extracting Reasoning Structures

Beyond assessing each cognitive element's presence, our span-level annotations enable analysis of how elements are temporally and hierarchically organized within traces. We encode each reasoning trace t as a heterogeneous transition graph G, where nodes represent elements from our 28-element taxonomy and edges capture one of three relationship types: CONTAINS (hierarchical nesting), NEXT (sequential ordering), or PARALLEL (co-occurrence). Each node and edge is weighted by its normalized pointwise mutual information (NPMI) with trace success.

Graph construction. We sort all annotated spans within trace t by start position, breaking ties by span length (descending). For each element pair (b_a, b_b) with spans $[s_a, e_a]$ and $[s_b, e_b]$ where $s_a \leq s_b$, we determine the relationship type through multi-stage classification:

1. PARALLEL: The Manhattan distance $|s_b - s_a| + |e_b - e_a|$ falls below threshold τ_{par} (elements co-occur).
2. CONTAINS: If $s_b \leq e_a \leq e_b$, we compute overlap ratio $\rho = (e_a - s_b)/(e_b - s_b)$; when $\rho > \tau_{\text{overlap}}$, b_a hierarchically encompasses b_b.
3. NEXT: If $e_a < s_b$ and overlap is below τ_{overlap}, or if spans partially overlap but ratio $(e_b - s_b)/(e_a - s_a) < \tau_{\text{overlap}}$, then b_a sequentially precedes b_b.

We set $\tau_{\text{par}} = 20$ characters and $\tau_{\text{overlap}} = 0.8$. To ensure sparsity and capture only direct dependencies, we retain only the first NEXT edge from each element to an immediately subsequent non-overlapping element. This yields a directed graph preserving both hierarchical decomposition and temporal ordering.

Common structure extraction. For each problem type, we extract $\hat{G}$ representing commonly deployed patterns. We aggregate transition statistics across all graphs $\{G_t\}$, computing edge frequencies $f(b_{\text{curr}}, b_{\text{next}}, e_{\text{type}}) = |\{\text{traces containing edge}\}|/|\{\text{total traces}\}|$. Starting from the most frequent initial element $b_{\text{start}} = \arg\max_b P(b \text{ appears first})$, we construct $\hat{G}$ via greedy forward search: at each step, we select the outgoing edge with maximum frequency among unvisited targets, maintaining acyclicity through a visited set. The process continues until reaching $|V_{\max}|$ nodes or exhausting valid edges.

Successful structure extraction. To identify patterns characteristic of correct reasoning, we extract G^* by selecting high-NPMI transitions. We begin from $b_{\text{start}} = \arg\max_b \text{NPMI}(b, \text{success})$ among elements initiating successful traces. At each step, we select the outgoing edge with maximum NPMI among unvisited targets. The process terminates when reaching $|V_{\max}|$ nodes, exhausting valid edges, or when all remaining edges have non-positive NPMI. This yields G^*, representing behavioral patterns strongly associated with success.

4.2 Successful and Common Structures Diverge

Figure 3 compares successful versus common reasoning structures for Algorithmic (well-structured) and Diagnosis (ill-structured) problems, revealing systematic misalignment.

Common patterns include failure-correlated elements. For Algorithmic problems, the most common pattern includes elements with *negative* NPMI: self-awareness (-0.141) and backtracking (-0.050), indicating these frequently deployed behaviors actually correlate with failure. Successful traces instead begin with selective attention followed by logical coherence and sequential organization—a pattern that identifies relevant features before proceeding.

Models bypass deliberate scoping on ill-structured problems. The divergence intensifies for Diagnosis problems. Successful traces follow a deliberate scoping strategy: selective attention $\rightarrow$ sequential organization $\rightarrow$ knowledge alignment *before* engaging forward chaining. This structure first identifies relevant features and aligns with problem & domain constraints before constructing solutions. The common pattern bypasses this scoping phase entirely, imme-

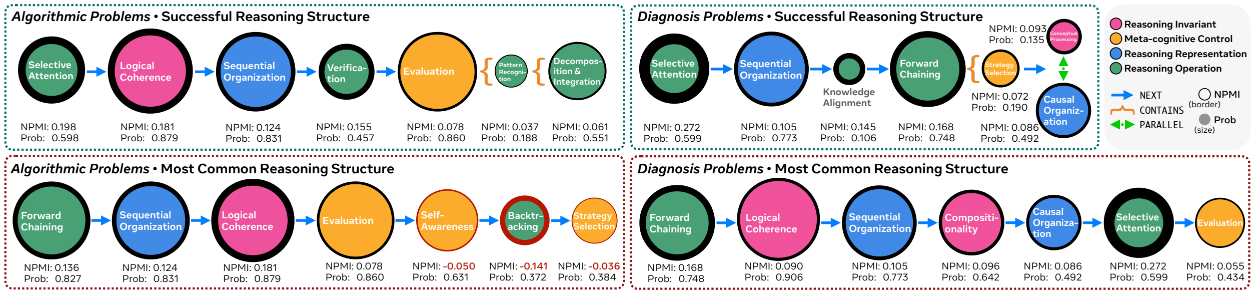

Figure 3: *Successful* (top) versus *common* (bottom) 7-node reasoning structures for Algorithmic and Diagnosis problems. Node size indicates NPMI with success; border thickness indicates transition probability. Common patterns include negative-NPMI elements; successful patterns show deliberate scoping before solution construction.

diately engaging forward chaining (probability 0.748). This premature solution-seeking explains systematic failures on problems requiring constraint satisfaction: models generate solutions before understanding what makes them valid.

Successful structures exhibit richer relationships. Across problem types, successful reasoning structures show more diverse structural relationships among cognitive elements (hierarchical nesting, parallel co-occurrence), whereas common traces connect elements almost exclusively through sequential transitions.

4.3 Test-Time Reasoning Guidance

We operationalize these findings by converting successful reasoning structures into test-time scaffolds that guide models toward effective cognitive patterns.

Methodology. For each problem type, we automatically convert the successful structure G^* into an actionable prompt. We linearize the graph structure and generate explicit scaffolding instructions for each cognitive element in sequence (e.g., "First, identify the relevant features of this problem [selective attention]... Then, organize your approach [sequential organization]... Next, align with domain knowledge [knowledge alignment]..."). This fully automated pipeline requires no expert prompt engineering, enabling assessment of whether models can leverage structural guidance without hand-crafted templates.

Evaluation. We evaluate on a stratified sample of approximately 50 problems per problem type, deliberately balanced between questions the model previously answered correctly and incorrectly. This design serves two purposes: (1) verifying that guidance does not degrade performance on already-solved problems, and (2) measuring improvement on previously failed instances. We quantify effectiveness as percentage change in accuracy relative to baseline.

4.4 Results

Table 3 presents performance improvements from cognitive structure guidance across 15 models and 10 problem types.

Substantial gains on ill-structured problems. Capable reasoning , particularly the Qwen3 family and larger R1-Distill variants, demonstrate significant improvements, with average gains reaching +26.7% on case analysis, +22.9% on diagnosis, and +21.4% on dilemmas. Individual improvements reach +72.0% (Olmo-3-7B on dilemmas), +66.7% (R1-Distill-Qwen-7B on diagnosis), and +60.0% (Qwen3-14B and R1-Distill-Qwen-32B on dilemmas). These gains concentrate on ill-structured problems where explicit structural scaffolding addresses the narrowing behavior identified in Section 3.2.

Capability threshold for guidance effectiveness. Effectiveness depends critically on model capacity. Smaller models, particularly DeepScaleR-1.5B and Hermes-3-8B, show pronounced *degradation*, with losses exceeding 50% on several problem types (e.g., -87.0% for DeepScaleR on story problems, -72.0% for both on algorithmic problems). This suggests a capability threshold: models must possess sufficient reasoning flexibility and instruction-following ability to productively incorporate structural guidance. Below this threshold, scaffolding could constrain rather than enhance reasoning, potentially by overwhelming limited cognitive resources or conflicting with ingrained heuristics.

Problem-type specificity mirrors analysis findings. The pattern of improvements directly echoes Figure 2: well-structured problems (algorithmic, story) show modest or negative effects across models, while ill-structured problems exhibit strong positive responses in capable systems. This alignment confirms that our intervention targets the identified misalignment by providing explicit organizational scaffolding for problems where models otherwise narrow to rigid sequential processing.

Table 3: Performance change (%) after test-time reasoning guidance: $\frac{\text{After}-\text{Before}}{\text{Before}} \times 100\%$. Approximately 50 problems per model-type pair, balanced between previously correct and incorrect. Color intensity indicates magnitude. Design problems showed catastrophic failure (0% accuracy) across all models and are excluded.

Model	Avg.	Log.	Algo.	Story	Rule	Dec.	Troub.	Diag.	Case	Dilem.	Fact.
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
Hermes-3-8B	-17.5	+0.0	-72.0	-68.0	-72.0	-4.0	-8.0	+41.2	+12.0	+56.0	+44.0
R1-Distill-Qwen-1.5B	-14.5	+8.5	-21.0	+33.3	+7.5	-56.0	-79.9	+0.0	+0.0	-23.9	-11.9
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
Olmo-3-7B	-3.5	-18.2	-25.1	-38.4	-12.0	+0.0	-36.0	+11.9	+12.0	+72.0	+0.0
OpenThinker-32B	-1.3	-9.4	-12.0	-30.0	+28.0	+12.0	+4.0	+6.9	+29.6	+37.5	+12.0
Qwen3-8B	+13.8	+3.1	+36.0	+11.1	+24.1	+16.0	+28.0	+48.0	+36.0	+7.9	+48.0
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
Qwen3-32B	+14.8	+11.8	+24.0	+20.6	+20.0	+24.0	+15.2	+24.2	+41.9	+19.5	+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-Qwen-32B	+22.3	+15.6	+36.0	+12.5	+40.0	+20.0	+32.0	+36.0	+56.0	+60.0	+40.0
Qwen3-14B	+22.5	+16.2	+32.0	+7.5	+44.0	+20.0	+44.0	+50.0	+52.0	+60.0	+32.0
Average	+2.0	+3.3	+1.1	+2.4	+12.9	+1.9	+3.6	+22.9	+26.7	+21.4	+16.3

Confirmation of latent capabilities. These results confirm that models possess cognitive capabilities they fail to deploy spontaneously. The same models that narrow to sequential processing on ill-structured problems successfully execute diverse reasoning strategies when explicitly scaffolded. Understanding cognitive behavioral patterns thus enables more effective model interaction. This suggests that training procedures inducing appropriate reasoning structures may be a promising direction for improving reasoning robustness.

5 Opportunities and Challenges

Our findings reveal that models possess cognitive capabilities they fail to deploy appropriately—narrowing to rigid sequential processing precisely where success demands diverse representations and meta-cognitive monitoring. This diagnosis suggests specific research directions. Current RL training rewards answer correctness without examining reasoning processes, creating systematic misalignment: models deploy rich behavioral repertoires where diversity is unnecessary while defaulting to inflexible strategies where adaptability is critical. Process reward models (Lightman et al., 2023; Uesato et al., 2022) address intermediate correctness but not behavioral diversity. Reward shaping targeting cognitive structure—not just accuracy—may better align training with behavioral patterns that predict success. Similarly, schema-based transfer theory (Gentner, 1983; Gick & Holyoak, 1980) explains why models succeed in-distribution but fail on superficial variants: transfer requires abstract structural representations, not surface-feature encoding. Training procedures encouraging structural comparison across problem types and diverse representational deployment may produce more robust generalization.

Our analysis also reveals a measurement gap: high-presence, low-PPMI elements (e.g., logical coherence at 91% presence but 0.091 PPMI) indicate models *attempt* behaviors without successfully *executing* them. Behavioral observation alone cannot distinguish genuine capabilities from pattern matching. Validation frameworks must test cognitive signatures: systematic transfer, perturbation robustness, and compositional deployment (Fodor & Pylyshyn, 1988). The relationship with cognitive science is bidirectional: models provide tools for testing cognitive theories at scale through manipulation impossible with human subjects, while human-model divergences—such as humans deploying meta-cognitive monitoring that models lack—constrain theories of what enables generalization beyond immediate task success. Our taxonomy provides shared vocabulary connecting cognitive science's theories with machine learning's implementations testable at scale.

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.

Paul Bloom and Lori Markson. Capacities underlying word learning. *Trends in cognitive sciences*, 2(2):67–73, 1998.

Matthew M Botvinick, Yael Niv, and Andew 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.

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, Junehyuk 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.

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.

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, Zhuoshu 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, Wanjia 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.

Evernote. AI Transcribe by Evernote, 2025. 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.

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.

Thomas L Griffiths, Falk Lieder, and Noah D Goodman. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, 7(2):217–229, 2015.

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.

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, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, 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.

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.

William James. The principles of psychology. *Henry Holt*, 1890.

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.

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/.

Charles Kemp and Joshua B Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, 2008.

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.

Wolfgang Kohler. *The mentality of apes*. Routledge, 2018.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi 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.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. URL https://arxiv.org/abs/2411.15124.

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

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.

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, 2025. URL https://arxiv.org/abs/2507.16746.

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, 2025.

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.

Marina Mancoridis, Bec Weeks, Keyon Vafa, and Sendhil Mullainathan. Potemkin understanding in large language models. *arXiv preprint arXiv:2506.21521*, 2025.

David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018.

Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, Nicholas Meade, Dongchan Shin, Amirhossein Kazemnejad, Gaurav Kamath, Marius Mosbach, Karolina Stańczak, and Siva Reddy. Deepseek-r1 thoughtology: Let's think about llm reasoning, 2026. URL https://arxiv.org/abs/2504.07128.

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.

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, 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.

George A Miller, Galanter Eugene, and Karl H Pribram. Plans and the structure of behaviour. In *Systems research for behavioral science*, pp. 369–382. Routledge, 2017.

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

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.

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.

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.

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

Team Olmo, :, 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, Lester James V. 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, 2025. URL https://arxiv.org/abs/2512.13961.

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.

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.

Jake Quilty-Dunn, Nicolas Porot, and Eric Mandelbaum. The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, 46:e261, 2023.

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

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

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

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

Oliver G Selfridge. Pandemonium: A paradigm for learning. In *Neurocomputing: foundations of research*, pp. 115–122. 1988.

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.

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.

Burrhus Frederic Skinner. *Science and human behavior*. Number 92904. Simon and Schuster, 1965.

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

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

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.

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.

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.

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, Yinger 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.

Dan Zakay and Richard A Block. Temporal cognition. *Current directions in psychological science*, 6(1):12–16, 1997.

A Appendix

A.1 Dataset and Model Specifications

A.1.1 Model Details

Table 4 provides complete specifications for all models and datasets analyzed in this work.

Table 4: 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
			Olmo 3 (7B, 32B) (Olmo et al., 2025)	2	21,224
	Total (Text Data)	**10,612**	**Total (Text Models)**	**16**	**169,792**
Audio	BLAB (Ahia et al., 2025)	417	Qwen3-Omni-30B-A3B-Thinking (Yang et al., 2025)	1	4,917
	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., 2025)	18,000	GPT 4.1 and Gemini 2.5 Pro (used for refinement, collapsed to 1)	1	18,000
	Total (Image Data)	**18,000**	**Total (Image Models)**	**1**	**18,000**
Grand Total (All Data)		**33,529**	**Grand Total (All Models)**	**18**	**192,709**

A.1.2 Text Reasoning Models

We analyze 16 open-weight text reasoning models spanning multiple architecture families and training paradigms:

Qwen3 Family. Qwen3 models (8B, 14B, 32B) feature hybrid thinking mode that enables extended reasoning (Yang et al., 2025). These models underwent comprehensive 4-stage reinforcement learning training including cold-start supervised fine-tuning, reasoning RL with GRPO, thinking mode fusion, and general RL for the 32B flagship, followed by efficient strong-to-weak distillation for smaller variants.

DeepSeek-R1 and Distillations. DeepSeek-R1 (671B parameters, 37B activated) serves as the frontier reasoning model and teacher for distilled variants (DeepSeek-AI et al., 2025). We analyze the full R1 model using released traces from GeneralThought due to compute constraints, plus distillations across two base architectures: Qwen2.5 (1.5B, 7B, 14B, 32B) and Llama3 (8B, 70B). Distillations were trained via supervised fine-tuning on 800K teacher-generated examples.

Olmo-3 Family. Olmo-3 models (7B, 32B) with thinking mode represent open development of reasoning capabilities (Olmo et al., 2025).

Community Reasoning Models. We include several community-developed reasoning-tuned models:

- **OpenThinker-32B** (Guha et al., 2025): Trained on 114K verified examples with automated verification (code execution for programming, LLM judge for mathematics), demonstrating that data quality can compensate for quantity.
- **s1.1-32B** (Muennighoff et al., 2025): Qwen-based model trained with efficient test-time scaling methods.
- **DeepScaleR-1.5B-Preview** (Luo et al., 2025): Compact reasoning model trained with scaled RL.
- **DeepHermes-3-Llama-3-8B-Preview** (Teknium et al., 2025): Hybrid reasoning model with user-controlled reasoning depth.

A.1.3 Text Reasoning Datasets

GeneralThought. The primary text reasoning dataset comprises 10,322 problems from GeneralThought (Taylor, 2024), which provides diverse tasks across domains. We down-sample extremely common task types (e.g., arithmetic, simple logic) using the dataset's task labels to maintain balanced representation across problem types. The dataset primarily focuses on verifiable tasks.

ClaimSpect. To supplement problem types underrepresented in verifiable benchmarks—particularly dilemmas requiring resolution of contradictory positions—we convert 290 real-world biomedical and geopolitical "nuanced" claims from ClaimSpect (Kargupta et al., 2025) into reasoning problems.

A.1.4 AUDIO REASONING

We evaluate Qwen3-Omni-30B-A3B-Thinking (Yang et al., 2025), the only open-weight audio-language model with thinking mode, on 4,917 problems from three datasets:

- **BLAB** (Ahia et al., 2025): 417 problems for long-form audio reasoning
- **MMAR** (Ma et al., 2025): 888 problems for diverse task coverage
- **MMAU-Pro** (Kumar et al., 2025): 3,612 problems for diverse skill coverage

We exclude commercial models (Gemini, GPT-4) because they produce summarized traces insufficiently reflective of underlying reasoning processes.

A.1.5 IMAGE REASONING

For image reasoning, we analyze 18,000 curated traces from Zebra-CoT (Li et al., 2025), selected for comprehensive task type coverage and high-quality reasoning traces. Real-world traces in Zebra-CoT 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-Pro and GPT-4.1) to refine them into diverse and coherent reasoning traces. We sample 1,000 question-reasoning pairs from each task type to obtain 18,000 traces. This dataset additionally enables analysis of cognitive elements in synthetic training data beyond raw model outputs.

A.1.6 HUMAN REASONING DATA

We collected 54 human reasoning traces as qualitative reference points for comparison with LLM-generated reasoning. We recruited 18 participants to solve a subset of GeneralThought problems while recording their reasoning using a think-aloud protocol. Participants verbalized their reasoning, which was later transcribed using Evernote AI Transcribe (Evernote, 2025).

Since some tasks require domain-specific facts or state tracking, participants were permitted to use tools including web search and note-taking. In such cases, participants verbalized tool usage (e.g., speaking search keywords aloud). Each trace was annotated separately by two human annotators using our 28-element taxonomy with a three-level scoring rubric (0=absent, 1=partially present, 2=present). Scores were aggregated via min-pooling to ensure conservative estimates.

These human traces overlap with a subset of the LLM evaluation set and illustrate how cognitive elements manifest in natural human reasoning, rather than establishing a full human benchmark. The annotation data were used for iteratively refining automatic span annotation prompts.

A.1.7 GENERATION PARAMETERS

For all model-generated reasoning traces (except DeepSeek-R1, for which we use released traces), we generate responses with temperature 0.6 and top-p 0.95 to balance diversity with coherence.

A.2 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.3 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.3.1 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.3.2 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.3.3 Classification Methodology

Each problem is classified through majority voting across three frontier LLMs (`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.3.4 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 192,709 traces.

A.4 Accuracy Analysis

Table 5 presents the complete accuracy results for all 16 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).

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%**

A.4.1 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.

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 $\approx$ 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 $\approx$ 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 $\approx$ 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.